# IMPROVED SEMI-PARAMETRIC TIME SERIES MODELS OF AIR POLLUTION AND MORTALITY

*Francesca Dominici, Aidan McDermott, Trevor J. Hastie*

May 16, 2004

## Abstract

In 2002, methodological issues around time series analyses of air pollution and health attracted the attention of the scientific community, policy makers, the press, and the diverse stakeholders concerned with air pollution. As the Environmental Protection Agency (EPA) was finalizing its most recent review of epidemiological evidence on particulate matter air pollution (PM), statisticians and epidemiologists found that the S-Plus implementation of Generalized Additive Models (GAM) can overestimate effects of air pollution and understate statistical uncertainty in time series studies of air pollution and health. This discovery delayed the completion of the PM Criteria Document prepared as part of the review of the U.S. National Ambient Air Quality Standard (NAAQS), as the time-series findings were a critical component of the evidence. In addition, it raised concerns about the adequacy of current model formulations and their software implementations.

In this paper we provide improvements in semi-parametric regression directly relevant to risk estimation in time series studies of air pollution. First, we introduce a closed form estimate of the asymptotically exact covariance matrix of the linear component of a GAM. To ease the implementation of these calculations, we develop the S package `gam.exact`, an extended version of `gam`. Use of `gam.exact` allows a more robust assessment of the statistical uncertainty of the estimated pollution coefficients. Second, we develop a bandwidth selection method to reduce confounding bias in the pollution-mortality relationship due to unmeasured time-varying factors such as season and influenza epidemics. Third, we introduce a conceptual framework to fully explore the sensitivity of the air pollution risk estimates to model choice. We apply our methods to data of the National Mortality Morbidity Air Pollution Study (NMMAPS), which includes time series data from the 90 largest US cities for the period 1987-1994.

**Key Words:** Semiparametric regression, time series, Particulate Matter (PM), Generalized Additive Model, Generalized Linear Model, Mean Squared Error, Bandwidth Selection.

**Affiliations:** Francesca Dominici, Associate Professor, Department of Biostatistics, Johns Hopkins University, Baltimore MD 21205; Aidan McDermott, Assistant Scientist, Department of Biostatistics Johns Hopkins University, Baltimore MD 21205; Trevor Hastie, Professor, Department of Statistics, Stanford University Palo Alto CA 94305-4065.

**Contact Information:** Francesca Dominici, e-mail: fdominic@jhsph.edu, phone: 410-6145107, fax: 410-9550958.

# 1 Introduction

Estimation of adverse health effects associated with ambient exposure to Particulate Matter (PM) constitutes one of the most interesting, recent case studies on the use of epidemiological evidence in public policy (Samet, 2000; Greenbaum et al., 2001). Under the Clean Air Act (Environmental Protection Agency, 1970), the US Environmental Protection Agency (EPA), is required: 1) to set National Ambient Air Quality Standards (NAAQS) for six “criteria” air pollutants at a level that protects the public’s health (Environmental Protection Agency, 1996, 2001), and 2) to periodically review these standards in light of the accumulated scientific evidence.

The periodic re-assessment of epidemiological evidence on the health effects of PM – which requires balancing a series of health effects, including hospitalization and death, against the feasibility and costs of further controls – creates a very sensitive social and political context. Estimates of the health effects of exposure to ambient PM and associated sources of uncertainty are at the center of an intense national debate, that has led to a high profile research agenda (National Research Council, 1998, 1999, 2001).

In the United States and elsewhere, evidence from time series studies of air pollution and health has been central to the regulatory policy process. Time series studies estimate associations between day-to-day variations in air pollution concentrations and day-to-day variations in adverse health outcomes, contributing epidemiological evidence useful for evaluating the risks of current levels of air pollution (Clancy et al., 2002; Lee et al., 2002; Stieb et al., 2002; Goldberg et al., 2003). Multi-site time series studies, like the National Morbidity Mortality Air Pollution Study (NMMAPS) (Samet et al., 2000a,c,b; Dominici et al., 2000, 2003), and the Air Pollution and Health: A European Approach (APHEA) study (Katsouyanni et al., 1997; Touloumi et al., 1997; Katsouyanni et al., 2001; Aga et al., 2003) which collected time series data on mortality, pollution, and weather in several locations in US and Europe, have been a key part of the evidence about the short-term effects of PM.

The nature and characteristics of time series data make risk estimation challenging, requiring complex statistical methods sufficiently sensitive to detect effects that can be small relative to the combined effect of other time-varying covariates. More specifically, the association between air pollution and mortality/morbidity can be confounded by weather and by seasonal fluctuations in health outcomes due to influenza epidemics, and to other unmeasured and slowly-varying factors (Schwartz et al., 1996; Katsouyanni et al., 1996; Samet et al., 1997). One widely used approach for a time series analysis of air pollution and health involves a semi-parametric Poisson regression with daily mortality or morbidity counts as the outcome, linear terms measuring the percentage increase in the mortality/morbidity associated with elevations in air pollution levels (the relative rates $\beta$s), and smooth functions of time and weather variables to adjust for the time-varying confounders.

In the last 10 years, many advances have been made in the statistical modelling of time series data on air pollution and health. Standard regression methods used initially have been almost fully replaced by semi-parametric approaches (Speckman, 1988; Hastie and Tibshirani, 1990; Green and Silverman, 1994) such as Generalized linear models (GLM) with regression splines (McCullagh and Nelder, 1989), Generalized additive models (GAM) with non-parametric splines (Hastie and Tibshirani, 1990) and GAM with penalized splines (Marx and Eilers, 1998). During the last few years, GAM with non-parametric splines was preferred to fully parametric formulations because of the increased flexibility in estimating the smooth component of the model, and the number of parameters to be estimated.

In 2002, as the Environmental Protection Agency (EPA) was finalizing its review of the evidence on particulate air pollution, statisticians found that the S implementation of GAM for time series analyses of air pollution and health can overestimate the air pollution effects and understate statistical uncertainty. More specifically, in these applications, the original default parameters of the `gam` function in S were found inadequate to guarantee the convergence of the backfitting algorithm

(Dominici et al., 2002b). In addition, the S function `gam`, in calculating the standard errors of the linear terms (the air pollution coefficients), approximates the smooth terms with linear functions, resulting in an underestimation of uncertainty (Chambers and Hastie, 1992; Ramsay et al., 2003; Klein et al., 2002; Lumley and Sheppard, 2003; Samet et al., 2003).

Computational and methodological concerns in the GAM implementation for time series analyses of pollution and health delayed the review of the National Ambient Air Quality Standard (NAAQS) for PM, as the time series findings were a critical component of the evidence. The EPA deemed it necessary to re-evaluate all of the time series analyses that used GAM and were key in the regulatory process. EPA officials identified nearly 40 published original articles and requested that the investigators reanalyze their data using alternative methods to GAM. The re-analyses were peer reviewed by a special panel of epidemiologists and statisticians appointed by the Health Effects Institute (HEI). Results of the re-analyses and a commentary by the special panel have been published in a Special Report of HEI (The HEI Review Panels, 2003; Dominici et al., 2003; Schwartz et al., 2003).

Recent re-analyses of time series studies have highlighted a second important epidemiological and statistical issue known as confounding bias. Pollution relative rate estimates for mortality/ morbidity could be confounded by observed and unobserved time-varying confounders (such as weather variables, season, and influenza epidemics) that vary in a similar manner as the air pollution and mortality/morbidity time series. To control for confounding bias, smooth functions of time and temperature variables are included into the semi-parametric Poisson regression model.

Adjusting for confounding bias is a more complicated issue than properly estimating the standard errors of the air pollution coefficients. The degree of adjustment for confounding factors, which is controlled by the number of degrees of freedom in the smooth functions of time and temperature ($df$), can have a large impact on the magnitude and statistical uncertainty of the mortality/morbidity relative rate estimates. In the absence of strong biological hypotheses, the choice of *df* has been based on expert judgment (Kelsall et al., 1997; Dominici et al., 2000), or on optimality criteria, such as minimum prediction error (based on the Akaike Information Criteria) and/or minimum sum of the absolute value of the partial autocorrelation function of the residuals (Touloumi et al., 1997; Burnett et al., 2001).

Motivated by these arguments, in this paper we provide the following computational and methodological contributions in semi-parametric regression directly relevant to risk estimation in time series studies of air pollution and mortality.

- We calculate a closed form estimate of the asymptotically exact covariance matrix of the linear component of a GAM (the air pollution coefficients). Furthermore, we developed the S package `gam.exact`, an extended version of `gam`, that implements these estimates. Hence `gam.exact` improves estimation of the statistical uncertainty of the air pollution risk estimates.

- We calculate the asymptotic bias and variance of the air pollution risk estimates as we vary the number of degrees of freedom in the smooth functions of time and temperature. Based upon these calculations, we develop a bandwidth selection strategy for the smooth functions of time and temperature that leads to air pollution risk estimates with small confounding bias with respect to their standard error. We apply the bandwidth selection method to four NMMAPS cities with daily air pollution data.

- We illustrate a statistical approach that allows a transparent exploration of the sensitivity of the air pollution risk estimates to degree of adjustment for confounding factors and more in general to model choice. Our approach is applied to data of the National Mortality Morbidity Air Pollution Study (NMMAPS), which includes time series data from the 90 largest US cities for the period 1987-1994.

By allowing a more robust assessment of all sources of uncertainty in air pollution risk estimates, including standard error estimation, confounding bias, and sensitivity to model choice, the application of our methods will enhance the credibility of time series studies in the current policy debate.

## 2 Statistical Model

Semi-parametric model specifications for time-series analyses of air pollution and health have been extensively discussed in the literature (Burnett and Krewski, 1994; Kelsall et al., 1997; Katsouyanni et al., 1997; Dominici et al., 2000; Zanobetti et al., 2000; Schwartz, 2000) and are briefly reviewed here. Data consist of daily mortality or morbidity counts ($y_t$), daily levels of one or more air pollution variables ($x_{1t}, \dots, x_{Jt}$), and additional time-varying covariates ($u_{1t} \dots, u_{Lt}$) to control for slow-varying confounding effects such as season and weather. Regression coefficients are estimated by assuming that the daily number of counts has an overdispersed Poisson distribution $E[Y_t] = \mu_t$, $\mathrm{Var}[Y_t] = \phi\mu_t$ and

$$\log \mu_t = \beta_0 + \sum_j \beta_j x_{jt} + \sum_{\ell=1}^{L} f_\ell(u_{\ell t}, d_\ell). \tag{1}$$

In our application, $\beta_j$ describes the percentage increase in mortality/morbidity per unit increases in ambient air pollution levels $x_{jt}$. The functions $f(\cdot, d_\ell)$ denote smooth functions of calendar time, temperature, and humidity, often constructed using smoothing splines, loess smoothers, or natural cubic splines with smoothing parameters $d_\ell$.

# 3 Asymptotically Exact Standard Errors in GAM

In this section we develop an explicit expression for the asymptotically exact (a.e.) statistical covariance matrix of the vector of the regression coefficients $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_J]$ corresponding to the linear component of model (1) when $f$ are modelled using smoothing splines and a GAM is used. Note that when $f$s are modelled using regression splines (such as natural cubic splines), model (1) becomes fully parametric and it is fitted by using Iteratively Re-weighted Least Squares (IRLS) (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989), and asymptotically exact standard errors are returned by the S-plus function `glm`.

An explicit expression for the a.e. covariance matrix of $\widehat{\boldsymbol{\beta}}$ can be obtained from the closed form solution for $\widehat{\boldsymbol{\beta}}$ from a backfitting algorithm (Hastie and Tibshirani, 1990, page 154):

$$\widehat{\boldsymbol{\beta}} \;=\; H\boldsymbol{z}, \text{ where } H = \left\{\boldsymbol{X}^t W(I - \boldsymbol{S})\boldsymbol{X}\right\}^{-1} \boldsymbol{X}^t W(I - \boldsymbol{S}),$$

and $\boldsymbol{X}$ is the $T \times J$ model matrix with columns $\boldsymbol{x}_j = [x_{j1}, \ldots, x_{jT}]^t$; $\boldsymbol{z}$ is the working response from the final iteration of the IRLS algorithm (McCullagh and Nelder, 1989) defined as $z_t = \hat{\eta}_t + (y_t - \hat{\mu}_t)/\hat{\mu}_t$; $W$ is diagonal in the final IRLS weights; and $\boldsymbol{S}$ is the $T \times T$ operator matrix that fits the additive model involving the smooth terms in the semi-parametric model (1). The total number of degrees of freedom in the smooth part of the model is defined as the trace of the additive operator matrix $\boldsymbol{S}$. Notice that here we have put all the additive smooth terms $\sum_{\ell=1}^{L} f_\ell(u_{\ell t}, d_\ell)$ together, and $\boldsymbol{S}$ represents the operator for computing this additive fit. As such, $\boldsymbol{S}$ represents a backfitting algorithm on just these terms.

From the definition of $\widehat{\boldsymbol{\beta}}$ above and the usual asymptotics we find that:

$$\widehat{\text{var}}(\widehat{\boldsymbol{\beta}}) \;=\; HW^{-1}H^t, \text{ where } W^{-1} = \widehat{\text{cov}}(\boldsymbol{z}).$$

Because calculation of the operator matrix $\boldsymbol{S}$ can be computationally expensive, the current version of the S-plus function `gam` approximates $\text{var}(\widehat{\boldsymbol{\beta}})$ by effectively assuming that the smooth component of the semi-parametric model is linear. That is, $\text{var}(\widehat{\boldsymbol{\beta}})$ is approximated by the appropriate submatrix of $(\boldsymbol{X}_{aug}^t W \boldsymbol{X}_{aug})^{-1}$, where $\boldsymbol{X}_{aug}$ is the model matrix of model (1) augmented by the predictors used in the smooth component of the model, i.e. $\boldsymbol{X}_{aug} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_L]^t$ (Hastie and Tibshirani, 1990; Chambers and Hastie, 1992).

In time series studies of air pollution and mortality, the assumption of linearity of the smooth component of model (1) is inadequate, resulting in underestimation of the standard error of the air pollution effects (Ramsay et al., 2003; Klein et al., 2002). The degree of underestimation tends to increase with the number of degrees of freedom used in the smoothing splines, because a larger number of non-linear terms is ignored in the calculations.

However, if $\boldsymbol{S}$ is a symmetric operator matrix, then $H$ can be re-defined as
$H = \left\{\boldsymbol{X}^t \left(W\boldsymbol{X} - W\boldsymbol{S}\boldsymbol{X}\right)\right\}^{-1} \left(W\boldsymbol{X} - W\boldsymbol{S}\boldsymbol{X}\right)^t$. Notice that symmetry in this case is with respect to a $W$ weighted inner product, and implies that $W\boldsymbol{S} = \boldsymbol{S}^t W$; weighted smoothing splines are symmetric, as are weighted additive model operators that use weighted smoothing splines as building blocks. Hence the expensive part of the calculation of $\widehat{\text{var}}(\widehat{\boldsymbol{\beta}})$ involves the calculation of the $T \times J$ matrix $\boldsymbol{S}\boldsymbol{X}$, having as column $j$ the fitted vector resulting from fitting the (weighted) additive model $\sum_{\ell=1}^{L} f_\ell(u_{\ell t}, d_\ell)$ to a "response" $\boldsymbol{x}_j$.

In summary, the calculations of $\boldsymbol{z}, W$ and $\boldsymbol{S}\boldsymbol{X}$ can be described in two steps: 1) fit model (1) using `gam` and extract the weights $\boldsymbol{w}$, as well as the actual degrees of freedom used in the backfitting $d_\ell^*$. Notice that the actual degrees of freedom may differ slightly from those requested in the call to `gam`, as a consequence of the changing weights in the IRLS algorithm. The weights $\boldsymbol{w}$ are the diagonal elements of the matrix $W$; 2) smooth each column of $\boldsymbol{X}$ with respect to $\sum_{\ell=1}^{L} f_\ell(u_{\ell t}, d_\ell^*)$, by using a `gam` with identity link and weights $\boldsymbol{w}$. The columns of $\boldsymbol{S}\boldsymbol{X}$ are the corresponding fitted values. Steps 1 and 2 are implemented in our S-plus function `gam.exact`, which returns the a.e. covariance matrix of $\widehat{\boldsymbol{\beta}}$ for any GAM. The software is available at `http://www.ihapss.jhsph.edu/software/gam.exact`.

For any smoother, the calculation of the variance of $\widehat{\boldsymbol{\beta}}$ requires the computation of $\boldsymbol{S}$. If $\boldsymbol{S}$ is symmetric, then we gain computational efficiency because we need to calculate $\boldsymbol{SX}$ only. If $\boldsymbol{S}$ is not symmetric, then we need to calculate $\boldsymbol{S}$ itself, which can be quite expensive for very long time series. Notice also that, because of the availability of a closed form solution of the back-fitting estimate of the smooth part of the GAM model — that is $\widehat{\boldsymbol{f}} = \boldsymbol{S}_f \boldsymbol{y}$, where $\boldsymbol{S}_f$ is the $T \times T$ smooth operator for $\boldsymbol{f}$ (Hastie and Tibshirani, 1990, page 127) — then our results can be also applied to calculate asymptotically exact confidence bands of $\widehat{\boldsymbol{f}}$, in addition to $\widehat{\boldsymbol{\beta}}$.

Finally, although we have detailed the standard error calculations for a semi-parametric model with log link and Poisson error, these calculations can be generalized for the entire class of link functions for GLM by calculating $z_t = \hat{\eta}_t + (y_t - \hat{\mu}_t)\frac{\partial \hat{\eta}_t}{\partial \hat{\mu}_t}$ (Nelder and Wedderburn, 1972) in step 2. In the simpler case of a Gaussian regression, the asymptotic covariance matrix $\text{var}(\widehat{\boldsymbol{\beta}})$ can be obtained by setting $\boldsymbol{w} = 1$ and $z_t = y_t$. Details of these calculations in this case have been discussed by Durban et al. (1999).

# 4 Understanding bias in semi-parametric regression

In this section we show that in order to remove systematic bias in the pollution effects, it is sufficient to model the seasonal effects with only enough degrees of freedom to capture the dependence of the pollution variable on those seasonal variables. More specifically, our goal is to estimate the association between air pollution ($x_t$) and mortality ($y_t$), denoted by the parameter $\beta$, in presence of seasonally varying confounding factors such as weather and influenza epidemics. We assume that these time-varying factors might affect $y_t$ by a function $f(t)$, and they might affect $x_t$ by a function $g(t)$. Let $\widehat{\beta}_d$ be the estimate of the air pollution coefficient corresponding to $d$ degrees of freedom in the spline representation of $f(t)$. Our statistical/epidemiological target is to determine $d$ that reduces confounding bias of $\widehat{\beta}_d$ with respect to its standard error. In this section we calculate the asymptotic bias and variance of $\widehat{\beta}_d$ as we vary the complexity in the representation of $f(t)$ with respect to $g(t)$ and we provide a bootstrap-based procedure for selecting $d$.

We consider a simple additive model of the following form:

$$y_t = \beta x_t + f(t) + \epsilon_t,\ \epsilon_t \sim N(0, \sigma^2),\ \sigma^2 > 0 \tag{2}$$

and we assume that the dependence between $x_t$ and $t$ is described by

$$x_t = g(t) + \xi_t,\ \xi_t \sim N(0, \sigma_\xi^2),\ \sigma_\xi^2 > 0. \tag{3}$$

We then represent $f(t)$ by a basis expansion $f(t) = \sum_{\ell=1}^{r} h_\ell(t)\delta_\ell$ or in vector notation $f(t) = \boldsymbol{h}^t(t)\boldsymbol{\delta}$. For a given set of $T$ time points, we can represent the vector of function values by $\boldsymbol{f} = H\boldsymbol{\delta}$, where $H$ is a $T \times r$ basis matrix. Without loss of generality we assume that $H^tH = TI$. We are therefore assuming that the $h_\ell(t)$ are mutually orthogonal, and are size-standardized. The factor $T$ is needed in asymptotic arguments below, and is realistic in the following sense. Suppose that $f$, and hence each of the $h_\ell$, are periodic (with a period of a year). We standardize them so that $\int_{\text{Year}} h_l^2(t)dt = 1$, or $\sum_{t=1}^{365} h_\ell(t)^2/365 = 1$. Then the sum-of-squares over $m$ years of data will be $T = 365 \cdot m$.

We start by assuming that $g(t)$ is smoother than $f(t)$, that is we assume that $g(t) = \boldsymbol{h}_1^t(t)\boldsymbol{\gamma}$, where $\boldsymbol{h}_1(t)$ is a subset of $q < r$ of the basis functions in $\boldsymbol{h}(t) = (\boldsymbol{h}_1(t), \boldsymbol{h}_2(t))$. Note that here $q$ and $r$ represent the number of degrees of freedom in the spline representations of $g(t)$ and $f(t)$, respectively. Simple calculations show that, if we model $f(t)$ by using enough basis functions to fully represent the relationship between $x_t$ and $t$ (i.e. $\hat{f}(t) = \sum_{\ell=1}^{q} h_\ell(t)\hat{\delta}_\ell = \boldsymbol{h}_1(t)\hat{\boldsymbol{\delta}}_1$), then:

$$\begin{aligned} \text{Bias}(\widehat{\beta}_q \mid \boldsymbol{x}) &= \boldsymbol{\xi}^t H_2 \boldsymbol{\delta}_2 / \left[\boldsymbol{\xi}^t (I - H_1 H_1^t/T)\boldsymbol{\xi}\right], \text{ and} \\ \text{Var}(\widehat{\beta}_q \mid \boldsymbol{x}) &= \sigma^2 / \left[\boldsymbol{\xi}^t (I - H_1 H_1^t/T)\boldsymbol{\xi}\right]. \end{aligned}$$

The denominator of $\text{Var}(\widehat{\beta}_q \mid \boldsymbol{x})$ is distributed as $\sigma_\xi^2 \chi_{T-q}^2$ with mean value $\sigma_\xi^2(T - q)$. It can be easily showed that squared bias and variance are both asymptotically negligible at rate $O_p(1/T)$ (see Appendix for details). Note that as we increase the number of basis functions in the representation of $f(t)$ (larger $q$) the bias diminishes (is zero for $q = r$) and the variance increases.

We now assume that $g(t)$ is more wiggly than $f(t)$, that is $g(t) = \boldsymbol{h}^t(t)\boldsymbol{\gamma}$ and that $f(t) = \boldsymbol{h}_1^t(t)\boldsymbol{\delta}$. As in the previous case, simple calculations show that if we model $f(t)$ with enough basis functions to adequately represent the relationship between $x_t$ and $t$ (i.e. $\hat{f}(t) = \sum_{\ell=1}^{r} h_\ell(t)\hat{\delta}_\ell = \boldsymbol{h}(t)\hat{\boldsymbol{\delta}}$), then:

$$\begin{aligned} \text{Bias}(\widehat{\beta}_r \mid \boldsymbol{x}) &= 0, \text{ and} \\ \text{Var}(\widehat{\beta}_r \mid \boldsymbol{x}) &= \sigma^2 / \left[\boldsymbol{\xi}^t(I - HH^t/T)\boldsymbol{\xi}\right]. \end{aligned} .$$

The denominator of $\text{Var}(\widehat{\beta}_r \mid \boldsymbol{x})$ is distributed as $\sigma_\xi^2 \chi_{T-r}^2$ with mean value $\sigma_\xi^2(T - r)$. Notice that by modelling $\hat{f}(t)$ with $r$ basis functions, we include into the regression model for $y_t$ a larger number of basis functions than it would be needed under a true model. This leads to an unbiased estimate of $\widehat{\beta}_r$, although with an inflated statistical variance.

In summary our asymptotic results suggest that modelling $f(t)$ with enough degrees of freedom to represent the relationship between $x_t$ and $t$ adequately, leads to an asymptotically unbiased estimate of the air pollution coefficient. In addition, as we increase the complexity in the representation of $f(t)$, that is as $d$ increases, then the bias of $\widehat{\beta}_d$ decreases and its standard error increases.

We use these asymptotic results to develop a bootstrap analysis to identify $d$ that leads to an efficient estimate of $\widehat{\beta}_d$, under the assumption that the exact forms of $g(t)$ and $f(t)$ are unknown. The computational steps of our bootstrap analysis are described below:

1. estimate the number of degrees of freedom $\widehat{d}$ that best predict $x_t$ as function of $t$. Generalized cross-validation (GCV) methods (Hastie and Tibshirani, 1990; Hastie et al., 1993) can be used to estimate $\widehat{d}$;

2. our asymptotic analysis has shown that if $g(t)$ is smoother than $f(t)$ then $\widehat{\beta}_{\widehat{d}}$ is asymptotically unbiased, and if $g(t)$ is rougher than $f(t)$ then $\widehat{\beta}_{\widehat{d}}$ is unbiased. Therefore if we fit the model $y_t = \beta x_t + f(t) + \epsilon_t$ by representing $f(t)$ with a number of degrees of freedom larger than $\widehat{d}$, say $\widehat{d}^\star = K \times \widehat{d}$ with $K \geq 3$ then $\widehat{\beta}_{\widehat{d}^\star}$ is unbiased but it has a large variance;

3. we then implement the following bootstrap analysis for identifying a number of degrees of freedom smaller than $\widehat{d}^{\star}$ that will lead to an estimate of the air pollution coefficient more efficient than $\widehat{\beta}_{\widehat{d}^{\star}}$;

4. for each bootstrap iteration $b = 1, \ldots, B$:

   - sample $y_t^b$ from the fitted full model in 2. obtained by using $\widehat{d}^{\star}$ degrees of freedom;
   - for $d = 1, \ldots \widehat{d}, \ldots, \widehat{d}^{\star}$, estimate $\widehat{\beta}_d^b$ by fitting the model $y_t^b = \beta_d x_t + \sum_{\ell=1}^{d} h_\ell(t)\delta_\ell + \epsilon_t$;

5. calculate bias and variance of $\widehat{\beta}_d^b$ as function of $d$ and select $d$ that leads to an unbiased estimate with small variance.

The proofs of the asymptotic results are summarized in the Appendix.

Notice that the success of our method relies upon the hypothesis that $\sigma_\xi^2 > 0$, or in other words that the air pollution levels $x_t$ fluctuates around $g(t)$ with measurement error. In fact under extreme confounding where the $g(t)$ is perfectly correlated with $x_t$ (i.e. $\sigma_\xi^2 \simeq 0$), then the the parameter $\beta$ is not identifiable. See The HEI Review Panels (2003) for examples illustrating how other df-selection strategies like the AIC fail in presence of extreme confounding.

In addition, the results presented in this section assume that $f(t)$ and $g(t)$ are modelled by the use of orthogonal basis functions, as for example, regression splines. Similar results when $f(t)$ and $g(t)$ are modelled by use of kernel smoothers are discussed in Green et al. (1985) and Speckman (1988). For smoothing splines, the analysis is complicated by the fact that all components of functions $f(t)$ and $g(t)$ (apart from the linear components), are modelled with bias. These biases depend on the complexity (roughness) of the component and the $d$ used, and will disappear asymptotically if $d$ grows appropriately (Green and Silverman, 1994).

### 4.1 Simulation Study

We further illustrate the performance of our bootstrap analysis by the implementation of the following simulation study. We generate $N$ data sets $(x_t^i, y_t^i)$ with known parameters and known $f(t)$ and $g(t)$ having the following spline representations:

$$\begin{aligned} f(t) &= a_0 + \textstyle\sum_{\ell=1}^{m_1} a_\ell h_\ell(t) \\ g(t) &= b_0 + \textstyle\sum_{\ell=1}^{m_2} b_\ell h_\ell(t) \end{aligned} \tag{4}$$

where $h_\ell(t)$ are known orthonormal basis functions, and $m_1$ and $m_2$ are the number of degrees of freedom in the spline representations of $f(t)$ and $g(t)$, respectively. We consider the following two scenarios:

(A) $g(t)$ is more smooth than $f(t)$, and we set $\beta = 0,\ m_1 = 10, m_2 = 4, \sigma = 0.17, \sigma_\xi = 3$.

(B) $g(t)$ is more wiggly than $f(t)$, and we set $\beta = 0,\ m_1 = 4, m_2 = 10, \sigma = 0.17, \sigma_\xi = 3$.

We obtain the spline coefficients (the $a$s and $b$s) used to create the scenarios by fitting the models $Y_t = a_0 + \sum_{\ell=1}^{m_1} a_\ell h_\ell(t) + \epsilon_t$ and $x_t = b_0 + \sum_{\ell=1}^{m_2} b_\ell h_\ell(t) + \xi_t$ to the Minneapolis log-mortality and $PM_{10}$ levels, respectively. We chose values of $\sigma$ and $\sigma_\xi$ to reflect the estimated standard errors of the observed log-mortality time series and $PM_{10}$ levels in Pittsburgh 1987-1988 with respect to smooth functions of time with $m_1 = 10$ and $m_2 = 4$ degrees of freedom, respectively. For each simulated data set $(x_t^i, y_t^i),\ i = 1, \ldots, N$ we:

1. estimate $m_2$ so that $g(t)$ is well modelled in the spline representation to adequately predict $x_t$;

2. fit the model $y_t = \beta x_t + f(t) + \epsilon_t$ by representing $f(t)$ with $\widehat{m}_2$ basis functions and calculate $\hat{\beta}_{\widehat{m}_2}$. Our our asymptotic analysis has shown that if $g(t)$ is smoother than $f(t)$ then $\widehat{\beta}_{\widehat{m}_2}$ is

asymptotically unbiased, and if $g(t)$ is rougher than $f(t)$ then $\widehat{\beta}_{\widehat{m}_2}$ is unbiased. Therefore if we fit the model $y_t = \beta x_t + f(t) + \epsilon_t$ by representing $f(t)$ with a number of degrees of freedom larger than $\widehat{m}_2$, say $\widehat{m}_2^\star = K \times \widehat{m}_2$ with $K \geq 3$ then $\widehat{\beta}_{\widehat{m}_2^\star}$ is unbiased but it will have a large variance;

3. implement the boostrap analysis for identifying a number of degrees of freedom smaller than $\widehat{m}_2^\star$ that will lead to a more efficient estimate than $\widehat{\beta}_{\widehat{m}_2^\star}$;

4. for each bootstrap iteration $b = 1, \ldots, B$ we:

   - sample $y_t^b$ from the fitted full model in 4 obtained by using $\widehat{m}_2^\star$ degrees of freedom.
   - for $d = 1, 2, \ldots M$, estimate $\hat{\beta}_d^b$ by fitting the model $y_t^b = \beta_d x_t + \sum_{\ell=1}^{d} h_\ell(t)\delta_\ell + \epsilon_t$.

We then calculate: 1) the average of the bootstrap estimates $\hat{\beta}_d^{\bullet,i} = \frac{1}{B}\sum_{b=1}^{B} \hat{\beta}_d^{b,i}$; 2) the Unconditional Squared Bias (USB): $\text{USB}_d = \frac{1}{N}\sum_{i=1}^{N}(\hat{\beta}_d^{\bullet,i} - \hat{\beta}_{\widehat{m}_2^\star}^{i})^2$; and 3) the Unconditional Variance (UV): $\text{UV}_d = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}_d^{b,i} - \hat{\beta}_d^{\bullet,i})^2$.

Figure 1 shows the results of the simulation study when $g(t)$ is smoother than $f(t)$ (scenario A) and when $g(t)$ is more wiggly than $f(t)$ (scenario B), respectively. The first row shows the true $g(t)$ (solid line), the estimated $\widehat{g}(t)$ (dotted line), one realization of the pollution time series $x_t$. The estimated $\widehat{g}(t)$ is obtained by fitting the model $x_t = \sum_{\ell=1}^{\widehat{m}_2} \gamma_\ell h_\ell(t) + \xi_t$, where $\widehat{m}_2$ is the average across the $N$ data sets of the estimated degrees of freedom from `bruto`. The excellent agreement between the solid and the dotted lines, support the use of `bruto` as a good strategy for estimating $m_2$. The second row shows the boxplots of the $N$ estimates ($\hat{\beta}_d^{\bullet,i} = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_d^{b,i}$) as function of $d$. The dots are plotted in correspondence of the unconditional average standard errors $\sqrt{\text{UV}_d}$. Notice in both scenarios A and B, as $d$ increases bias decreases and standard error increases. The third row shows the unconditional squared bias ($\text{USB}_d$) (triangles) and the unconditional variance ($\text{UV}_d$) (dots) as function of $d$. Under scenario A, as $d$ becomes larger than 4 the squared bias is zero and it is dominated by the variance. Under scenario B, USB becomes smaller than UV for $d$ larger than 7 and fades away for $d$ larger than 10.

## 5 NMMAPS Data Analysis

In this section, we apply our methods to the NMMAPS data base which is comprised of daily time series of air pollution levels, weather variables, and mortality counts for the largest 90 cities in the US from 1987 to 1994. A full description of the NMMAPS data base is detailed in Samet et al. (2000b) and data are posted on the web site `http://www.ihapss.jhsph.edu`. First, we apply our bootstrap analysis for removing confounding bias to four NMMAPS cities with daily data available. Second, we extend modelling approaches in a hierarchical fashion, and we estimate national average air pollution effects as function of degrees of adjustment for confounding factors. Details of the two data analyses are below.

To apply the boostrap analysis to the four NMMAPS cities with daily data, we use the following simplified version of the NMMAPS core model (Dominici et al., 2000, 2002c) $E[Y_t] = \mu_t$, $\mathrm{Var}[Y_t] = \phi\mu_t$ and

$$\log \mu_t = \beta_0(\alpha) + \beta(\alpha) PM_{10t} + s_1(t, d_1 \times \alpha) + s_2(\mathrm{temp}_t, d_2 \times \alpha) \tag{5}$$

where $Y_t$ is the daily number of deaths, $\phi$ is the over-dispersion parameter, $PM_{10t}$ is the daily level of PM with a mass median in aerodynamic diameter less than 10 micrometers ($\mu m$), temp is the temperature, and $t = 1, \ldots, 365 \times 8$ days. We assume $\alpha$ to be 25 equally-spaced points between $1/K$ and $K$, and $s$ to be regression splines with a natural spline basis.

First within each city, we estimate $(\widehat{d_1}, \widehat{d_2})$ in the smooth functions of time and temperature that "best" predict $PM_{10}$. Here we use generalized cross-validation (GCV) methods (Hastie and Tibshirani, 1990; Hastie et al., 1993). Table 1 summarizes the results for the four cities: the estimated $(\widehat{d_1}, \widehat{d_2})$, and $\widehat{\beta}_{\widehat{d_1},\widehat{d_2}}$s which denote the relative rate estimates obtained by using $(\widehat{d_1}, \widehat{d_2})$

in the smooth functions of time and temperature in the model (5). Based upon our asymptotic analysis, $\widehat{\beta}_{\widehat{d}_1,\widehat{d}_2}$s are asymptotically unbiased. In Seattle we estimated larger $\widehat{d}$s than in the other cities indicating a more complex relationship between $PM_{10}$ and the time-varying confounders, thus suggesting that we need large $d$'s to remove confounding bias. In Table 1 are also summarized city-specific estimates and 95% confidence intervals of $\widehat{\beta}_{\widehat{d}_1^\star,\widehat{d}_2^\star}$ where with $\widehat{d}_1^\star = K \times \widehat{d}_1$ and $\widehat{d}_2^\star = K \times \widehat{d}_2$. In Pittsburgh, Chicago, and Minneapolis we choose $K = 3$. In Seattle, because $K$ multiplies very large $d$s we choose $K = 2$ to easy the computations. Note that $\widehat{\beta}_{\widehat{d}_1^\star,\widehat{d}_2^\star}$ are unbiased because they are obtained by using smooth functions of time and temperatures that are much more flexible than the ones needed to model the relationship between $PM_{10}$ and time and temperature.

To implement out bootstrap analysis, first we sample 500 mortality time series from the fitted model (5) with $\widehat{d}_1^\star$ and $\widehat{d}_2^\star$. Second, for each bootstrap sample we re-fit model (5) with $(\alpha \times \widehat{d}_1, \alpha \times \widehat{d}_2)$ degrees of freedom and $\alpha$ varying from $1/K$ and $K$. Figure 2 (left panels) shows boxplots of the bootstrap distributions of $\widehat{\beta}^b(\alpha),\ b = 1, \ldots, 500$ as function of $\alpha$. Solid and dotted horizontal lines are placed at $\widehat{\beta}_{\widehat{d}_1^\star,\widehat{d}_2^\star}$ and at 0, respectively.

The asymptotic analysis suggests that for $\alpha$ smaller than 1 the bias can be substantial because we are using $d$s smaller than $\widehat{d}_1, \widehat{d}_2$. For $\alpha = 1$, although the bias is asymptotically zero, for finite samples bias can still occurr. For $\alpha$ larger than 1, bias diminishes and we assume that it is zero for $\alpha = K$. These results are confirmed in the bootstrap analysis. In Pittsburgh, Chicago and Seattle the boxplots shows a little bias for $\alpha = 1$, whereas in Minneapolis the bias is zero for $\alpha = 1$. For $\alpha > 1$ bias diminishes and it is not necessary to use $\alpha = K$ to remove it completely. In fact in Pittsburgh, Chicago and Seattle the bias is trascurable for $\alpha$ equal to 1.6, 1.8 and 1.9, respectively.

We now extend our analysis to the entire NMMAPS data base. The implementation of our bootstrap-based methodology here is complicated because $PM_{10}$ is measured approximately every six days in most of the NMMAPS locations, however we can still extend the NMMAPS model in an hierarchical fashion and estimate national average air pollution effects as function of $\alpha$. We consider the following overdispersed Poisson semi-parametric model used in the NMMAPS analyses

$$\begin{aligned}\log E[Y_t^c] &= \text{age-specific intercepts} + \beta^c(\alpha) PM_{10t}^c + s(t, 7/\text{year} \times \alpha) + \\ &+ s(\text{temp}_t, 6 \times \alpha) + s(\text{dewpoint}_t, 3 \times \alpha) + \text{age} \times s(t, 8 \times \alpha)\end{aligned}$$

where $y_t^c$ is the daily number of deaths in city $c$, $PM_{10t}$ is the daily level of $PM_{10}$, temp and dew are the temperature and dewpoint temperature, and the age-specific intercepts correspond to the three age groups of younger than 65, between 65 and 75 and older than 75. Justification for the selection of the degrees of freedom to control for longer-term trends, seasonality and weather can be found in Samet et al. (1995,1997,2000a), Kelsall et al. (1997), and Dominici et al. (2000b).

Based upon the statistical analyses of the four cities with daily data and additional exploratory analyses, we set $\alpha$ to take on 25 equally spaced points varying from 1/3 to 3. As in the previous model formulation, this choice allows the degree of adjustment for confounding factors to vary greatly. We then assume the following two-stage normal-normal hierarchical model: Stage I) $\widehat{\beta}^c(\alpha) \sim N(\beta^c(\alpha), v^c(\alpha))$; Stage II) $\beta^\star(\alpha) \sim N(\beta^\star(\alpha), \tau^2(\alpha))$ where $\beta^\star(\alpha)$ and $\tau^2(\alpha)$ are the national average air pollution effects and the variance across cities of the true city-specific air pollution effects, both as a function of $\alpha$.

We fit the hierarchical model by using a Bayesian approach, with a flat prior on $\beta^\star(\alpha)$ and uniform prior on the shrinkage factor $\tau^2(\alpha)/\left[\tau^2(\alpha) + v^c(\alpha)\right]$ (Everson and Morris, 2000). Sensitivity of the national average estimates to the specification of the prior distribution of $\tau^2$ has been explored elsewhere (Dominici et al., 2002a).

To investigate sensitivity of the national average estimates to model choice, for each value of $\alpha$, we estimate $\widehat{\beta}^c(\alpha)$ and $v^c(\alpha)$ using three methods: 1) GAM with smoothing splines and approximated standard errors (GAM-approx s.e.); 2) GAM with smoothing splines and asymptotically exact standard errors (GAM-exact); and 3) GLM with natural cubic splines (GLM).

The left top panel of Figure 3 shows the national average estimates (posterior means) as a function of $\alpha$. Dots, octagons, and triangles denote estimates under GAM-approx s.e., GAM-exact, and GLM, respectively. The grey polygon represents 95% posterior intervals of the national average estimates under GAM-exact. The vertical segment is placed at $\alpha = 1$, that is, the degree of adjustment used in the NMMAPS model (Dominici et al., 2000). The black curves at the top right panel denote the city-specific Bayesian estimates of the relative rates under GAM-exact.

Figure 3 provides strong evidence for association between short-term exposure to $PM_{10}$ and mortality, which persists for different values of $\alpha$. Consistent with the results for the four cities, national average estimates decrease as $\alpha$ increase, and level off for $\alpha$ larger than 1.2 with a very modest increase in posterior variance. However even when $\alpha = 3$, the national average effect is estimated at 0.2% increase in total mortality for 10 $\mu g/m^3$ increase in $PM_{10}$ (95% posterior interval 0.05 to 0.35).

This picture also shows robustness of the results to model choice (GAM versus GLM). National average estimates under GAM-exact are slightly smaller than those obtained under GAM-approx, although this difference is very small. These two sets of estimates are comparable because in hierarchical models, underestimation of standard errors at the first stage ($\sqrt{v^c(\alpha)}$) is compensated by the overestimation of the heterogeneity parameter at the second stage ($\tau^2(\alpha)$). Thus the posterior total variance of the national average estimates remains approximately constant (Daniels et al., 2004).

The bottom left and right panels of Figure 3 show posterior means of the average s.e. of $\widehat{\beta}^c$ ($\sqrt{\frac{1}{90}\sum_c v^c(\alpha)}$), and of the heterogeneity parameters $\tau(\alpha)$. Because of the nature of the approximation, the average standard errors are smaller in GAM-approx than in GAM-exact or GLM, and do not vary with $\alpha$. If GAM-exact or GLM are used, then the average standard errors increase with $\alpha$, with GAM-exact providing slightly larger estimates. Under all three modelling approaches, the posterior mean of $\tau(\alpha)$ (heterogeneity) decreases as $\alpha$ increases, indicating that less control for confounding factors inflates the variability across cities of the $\beta^c(\alpha)$s.

# 6 Discussion

In this paper, we propose improvements in semi-parametric regression for time series analyses of air pollution and health. Our contributions are computational, methodological, and substantive. From a computational standpoint, we develop an algorithm for estimating the covariance matrix of the vector of the regression coefficients in GAM (the air pollution risk estimates) that properly accounts for the degree of adjustment for confounding factors. From a methodological standpoint, we calculate the asymptotic bias and variance of the air pollution risk estimate as we vary the degree of adjustment for confounding factors. We show that confounding bias can be removed by including in the Poisson regression model smooth functions of time and temperature that are flexible enough to predict pollution. For a substantive standpoint, we introduce a conceptual framework for exploring the sensitivity of the national average pollution effect as we vary the degree of adjustment for confounding bias and the choice of the statistical model.

Our S-plus function `gam.exact` returns an asymptotically exact covariance matrix of the regression coefficients corresponding to the linear component of a GAM, and it can be used for any number of linear predictors, smooth terms, link functions, and distribution errors. These calculations are computationally efficient because they simply require the fit of as many GAM as there are regression coefficients in the linear component of the model (in our case-study the number of pollutants included in the model) instead of calculating the $T \times T$ smoother operator $\boldsymbol{S}$ (in our case-study, $T$ is equal to 8 years of daily data, and therefore the computation of $\boldsymbol{S}$ would have been almost prohibitive). However, this computational efficiency can be obtained for symmetric smoothers only, as for example smoothing splines. Simulation studies suggest that these standard error calculations are adequate for non symmetric smoothers when a GAM with identity link is used (Durban et al., 1999). A similar conclusion may hold for any link function, although additional investigations are warranted.

Selecting the number of degrees of freedom in the smooth functions of time and temperature to reduce confounding bias in the relative rate estimates is a more challenging problem than standard error calculations. Our asymptotic calculations show that in most situations where the air pollution levels are associated with time-varying confounders plus some measurement error, we can effectively reduce confounding bias by: 1) estimating the number of degrees of freedom in the smooth functions of time and temperature that best predict pollution levels; and 2) use those degrees of freedom as a starting point for implementing a bootstrap analysis that allows us to calculate bias and variance of the estimated pollution effects as function of $df$. Visual inspection of the boxplots of bootstrap estimates of $\beta$ as function of the degrees of freedom are informative for identifying the $df$ that leads to an unbiased estimate with small variance.

Controlling for the potential confounding effects of "measured confounders" (such as weather variables) is a better identified problem than controlling for "unmeasured confounders" (such as the seasonal fluctuations in health outcomes that cannot be attributed to seasonal fluctuations in pollution). The bandwidth selection problem for removing the effect of measured confounders could be based on prior work on optimal smoothing for generalized semi-linear models (Carroll et al., 1997; Emond and Self, 1997).

Recent re-analyses have renewed interest in methodological aspects of time series studies of air pollution and health and are informing the NAAQS process for PM (Dominici et al., 2003; Schwartz et al., 2003; The HEI Review Panels, 2003). In the re-analyzed time series studies, the increase in daily total mortality due to $10\mu/m^3$ increase in $PM_{10}$ has been estimated to be on the order of 0.2% to 0.8%. The increase in deaths from cardiac or respiratory related causes can be 4 to 5 times as large. The NMMAPS modelling approach was developed with grounding in the biomedical literature on pollution, weather, and mortality (Samet et al., 1997, 1998). It can be extended to allow for: 1) integration of scientific knowledge about the physics and chemistry of the association between pollution and weather; 2) interactions between current and past levels of weather variables to better control for confounding effects of heat waves; and 3) lagged pollution effects. Physical relationships between pollution and weather are very complex, they tend to vary from city to city, and integrating such information into the statistical formulation could be very challenging. In addition, in most of the NMMAPS locations, $PM_{10}$ levels are available only every six days, thus limiting the implementation of distributed lag models.

The use of epidemiological evidence for policy purposes when biological evidence of harm is still accruing places a heavy weight on analytic methods. In this sensitive political context, a transparent and comprehensive assessment of all sources of uncertainty would greatly enhance the utilization of time series findings for regulatory policy. Methods proposed in this paper and their applications to the NMMAPS improve the estimation of statistical uncertainty of the estimated risks, introduce a diagnostic tool to reduce confounding bias, and illustrate a conceptual framework to explore the sensitivity of the relative rates estimates to the degree of adjustment for confounding factors and more in general to model choices.

## 7 Appendix: Proofs of the asymptotic results in section 4

**$\boldsymbol{g}$ is smoother than $\boldsymbol{f}$:** we assume:

$$\begin{aligned} y_t &= \beta x_t + f(t) + \epsilon_t \text{ with } \epsilon_t \sim N(0,\sigma^2) \\ \boldsymbol{f} &= H_1\boldsymbol{\delta}_1 + H_2\boldsymbol{\delta}_2 \\ \boldsymbol{Y} &= \boldsymbol{x}\beta + H_1\boldsymbol{\delta}_1 + H_2\boldsymbol{\delta}_2 + \boldsymbol{\epsilon} \\ x_t &= g(t) + \xi_t \text{ with } \xi_t \sim N(0,\sigma^2_\xi) \\ \boldsymbol{g} &= H_1\boldsymbol{\gamma} \end{aligned} \tag{6}$$

where $dim(\boldsymbol{x}) = T \times 1$, $dim(H_1) = T \times q$ and $dim(H_2) = T \times (r-q)$. We assume that $H^tH = T \cdot I$; we use $T$ rather than 1, so that we can think of the coefficients $\boldsymbol{\delta}_1$, $\boldsymbol{\delta}_2$ and $\boldsymbol{\gamma}$ as staying fixed as $T$ increases.

We model $f$ by using sufficient degrees of freedom to fully represent the relationship between $x_t$ and $t$. Therefore, we fit a linear regression model having $\boldsymbol{y}$ as outcome, $[\boldsymbol{x}, H_1]$ as predictors, and let $\boldsymbol{\theta}_q$ be the corresponding vector of regression coefficients. The OLS estimate of $\boldsymbol{\theta}_q$ is so defined:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_q &= \left(\tilde{X}^t\tilde{X}\right)^{-1}\tilde{X}^t\boldsymbol{y} \text{ where}\\ \tilde{X} &= [\boldsymbol{x}, H_1]\end{aligned}$$

We have that:

$$\begin{aligned}E[\widehat{\boldsymbol{\theta}}_q \mid \boldsymbol{x}] &= \begin{bmatrix}\beta\\ \boldsymbol{\delta}_1\end{bmatrix} + \left(\tilde{X}^t\tilde{X}\right)^{-1}\tilde{X}^t H_2\boldsymbol{\delta}_2\\ &= \begin{bmatrix}\beta\\ \boldsymbol{\delta}_1\end{bmatrix} + \left(\tilde{X}^t\tilde{X}\right)^{-1}\begin{bmatrix}\boldsymbol{x}^t H_2\boldsymbol{\delta}_2\\ \boldsymbol{o}\end{bmatrix}\end{aligned}$$

The first line follows from writing

$$\boldsymbol{Y} = \tilde{X}\begin{bmatrix}\beta\\ \boldsymbol{\delta}_1\end{bmatrix} + H_2\boldsymbol{\delta}_2 + \boldsymbol{\epsilon},$$

and the second from the orthogonality of $H_1$ and $H_2$. Let $\widehat{\beta}_q$ be the first element of the vector $\widehat{\boldsymbol{\theta}}_q$, therefore:

$$\begin{aligned}E[\widehat{\beta}_q \mid \boldsymbol{x}] &= \beta + \frac{\boldsymbol{x}H_2\boldsymbol{\delta}_2}{||\boldsymbol{x}-H_1H_1^tT^{-1}\boldsymbol{x}||^2}\\ &= \beta + \frac{\boldsymbol{\xi}H_2\boldsymbol{\delta}_2}{(\boldsymbol{\xi}^t(I-H_1H_1^t/T)\boldsymbol{\xi})}\\ \mathrm{V}[\widehat{\beta}_q \mid \boldsymbol{x}] &= \frac{\sigma^2}{||\boldsymbol{x}-H_1H_1^tT^{-1}\boldsymbol{x}||^2} = \frac{\sigma^2}{\boldsymbol{\xi}^t(I-H_1H_1^t/T)\boldsymbol{\xi}}\end{aligned}$$

In the first line, $1/||\boldsymbol{x} - H_1H_1^tT^{-1}\boldsymbol{x}||^2$ is the top left element of the partioned inverse of $\tilde{X}^t\tilde{X}$; the second line uses the orthogonality of $H_1$ and the residual projection operator $(I - H_1H_1^t/T)$. The same arguments apply to the third line, using the standard formula for the covariance matrix of the least squares fit $\mathrm{cov}(\widehat{\boldsymbol{\theta}}_q) = (\tilde{X}^t\tilde{X})^{-1}\sigma^2$. In summary, if $g(t)$ is smoother than $f(t)$, and if we represent $f(t)$ in model (1) with enough basis functions to represent $g(t)$ in model (2) adequately, then:

1. the bias of $\widehat{\beta}_q$ can be written as $z_1/z_2$ where unconditionally $z_1 \sim N(0, \sigma^2 \cdot T \cdot ||\boldsymbol{\delta}_2||^2)$ and $z_2 \sim \sigma^2_\xi \chi^2_{T-q}$. These two terms are not statistically independent, so the most we can say is that this term is $O_p(1/\sqrt{T})$.

2. the denominator of the variance of $\widehat{\beta}_q$ is unconditionally distributed as $\sigma^2_\xi \chi^2_{T-q}$. Hence the standard error of $\widehat{\beta}_q$ is also $O_p(1/\sqrt{T})$

***g* rougher than *f*:** We now repeat the same type of calculations under the assumption that $g(t)$ is rougher than $f(t)$. We assume:

$$\begin{aligned} y_t &= \beta x_t + f(t) + \epsilon_t,\ \epsilon_t \sim N(0, \sigma^2) \\ \boldsymbol{f} &= H_1 \boldsymbol{\delta}_1 + H_2 \boldsymbol{\delta}_2,\ \text{where } \boldsymbol{\delta}_2 = \boldsymbol{o} \\ \boldsymbol{Y} &= \boldsymbol{x}\beta + H_1 \boldsymbol{\delta}_1 + \boldsymbol{\epsilon} \\ x_t &= g(t) + \xi_t,\ \xi_t \sim N(0, \sigma^2_{xi}) \\ \boldsymbol{g} &= H_1 \boldsymbol{\gamma}_1 + H_2 \boldsymbol{\gamma}_2 \end{aligned} \tag{7}$$

As before, we model $f$ by using sufficient degrees of freedom to fully represent the relationship between $x_t$ and $t$. Therefore, we fit a linear regression model having $\boldsymbol{y}$ as outcome, $[\boldsymbol{x}, H_1, H_2]$ as predictors, and let $\boldsymbol{\theta}_r$ be the corresponding vector of regression coefficients. Notice that here we using more basis functions that we would need under the true model for $y_t$. The OLS estimate of $\boldsymbol{\theta}_r$ is given by

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_r &= \left(\tilde{X}^t \tilde{X}\right)^{-1} \tilde{X}^t \boldsymbol{Y} \text{ where} \\ \tilde{X} &= [\boldsymbol{x}, H_1, H_2] = [\boldsymbol{x}, H] \end{aligned}$$

Standard least squares calculus shows that

$$\begin{aligned} E[\widehat{\boldsymbol{\theta}}_r \mid \boldsymbol{x}] &= \left(\tilde{X}^t \tilde{X}\right)^{-1} \tilde{X}^t E[\boldsymbol{Y}] \\ &= \left(\tilde{X}^t \tilde{X}\right)^{-1} \tilde{X}^t [\beta \boldsymbol{x} + H_1 \boldsymbol{\delta}_1 + H_2 \boldsymbol{o}] \\ &= \begin{bmatrix} \beta \\ \boldsymbol{\delta}_1 \\ \boldsymbol{o} \end{bmatrix} \end{aligned}$$

Let $\widehat{\beta}_r$ be the first element of $\widehat{\boldsymbol{\theta}}_r$, therefore:

$$\begin{aligned} E[\widehat{\beta}_r \mid \boldsymbol{x}] &= \beta \\ \mathrm{V}[\widehat{\beta}_r \mid \boldsymbol{x}] &= \tfrac{\sigma^2}{||\boldsymbol{x}^t(I-HH^t/T)\boldsymbol{x}||^2} = \tfrac{\sigma^2}{\boldsymbol{\xi}^t(I-HH^t/T)\boldsymbol{\xi}} \end{aligned}$$

In summary, if $g(t)$ is more wiggly than $f(t)$, and if we represent $f(t)$ with enough basis functions to capture the relationship between $x_t$ and $t$ in model (2), then:

1. $\widehat{\beta}_r$ is unconditionally unbiased;
2. the denominator of the variance of $\widehat{\beta}_r$ is unconditionally distributed as $\sigma^2_\xi \chi^2_{T-r}$.

## Acknowledgments

Funding for Francesca Dominici was provided by a grant from the Health Effects Institute (Walter A. Rosenblith New Investigator Award), by NIEHS RO1 grant (ES012054-01), and by NIEHS Center in Urban Environmental Health (P30 ES 03819). Trevor Hastie was partially supported by grant DMS-0204162 from the National Science Foundation, and grant RO1-EB0011988-08 from the National Institutes of Health. We would like to thank Drs Scott L. Zeger, Jonathan M. Samet, Giovanni Parmigiani, and Jamie Robins for comments.

# References

Aga, E., Samoli, E., Touloumi, G., Anderson, H., Cadum, E., Forsberg, B., Goodman, P., Goren, A., Kotesovec, F., Kriz, B., Macarol-Hiti, M., Medina, S., Paldy, A., Schindler, C., Sunyer, J., Tittanen, P., Wojtyniak, B., Zmirou, D., Schwartz, J., and Katsouyanni, K. (2003). "Short-term effects of ambient particles on mortality in the elderly: results from 28 cities in the APHEA2 Project." *European Respiratory Journal Supplement*, 40, 28–33.

Burnett, R. and Krewski, D. (1994). "Air Pollution effects of hospital admission rates: A random effects modelling approach." *The Canadian Journal of Statistics*, 22, 441–458.

Burnett, R., Ma, R., Jerrett, M., Goldberg, M., Cakmak, S., Pope, A., and Krewski, D. (2001). "The spatial association between community air pollution and mortality: a new method of analyzing correlated geographic cohort data." *Environmental Health Perspectives*, 109, 375–380.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). "Generalized Partially Linear Single-index Models." *Journal of the American Statistical Association*, 92, 477–489.

Chambers, J. M. and Hastie, T. (1992). *Statistical Models in S*. Chapman and Hall, London.

Clancy, L., Goodman, P., Sinclair, H., and Dockery, D. (2002). "Effect of air-pollution control on death rates in Dublin, Ireland: an intervention study." *Lancet*, 360, 1210–1214.

Daniels, M., Dominici, F., and Zeger, S. (2004). "Underestimation of Standard Errors in Time Series Studies of Air Pollution and Mortality." *Epidemiology*, 15, 57–62.

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., Daniels, M., Zeger, S. L., and Samet, J. M. (2003). *A Special Report* *to the Health Effects Institute on the Revised Analyses of the NMMAPS II Data*. The Health Effects Institute, Cambridge, MA.

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.

— (2002c). "Airborne particulate matter and mortality: Time-scale effects in four US Cities." *American Journal of Epidemiology*, 157, 1053–1063.

Dominici, F., Samet, J. M., and Zeger, S. L. (2000). "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, with discussion*, 163, 263–302.

Durban, M., Hackett, C., and Currie, I. (1999). "Approximate Standard Errors in Semiparametric Models." *Biometrics*, 55, 699–703.

Emond, M. and Self, S. G. (1997). "An Efficient Estimator for the Generalized Semilinear Model." *Journal of the American Statistical Association*, 92, 1033–1040.

Environmental Protection Agency (1970). "The Clean Air Act (CAA); 42 U.S.C. s/s 7401 et seq. (1970) Clean Air Act and Amendments of 1970 (PL 91-604; 42 USC 1857h-7 et seq.; amended 1970." *US Environmental Protection Agency*.

Environmental Protection Agency (1996). "Review of the National Ambient Air Quality Standards for Particulate Matter: Policy Assessment of Scientific and Technical Information. OAQPS Staff Paper. Research Triangle Park, North Carolina, U.S. Government Printing Office." *Environmental Protection Agency*.

— (2001). "Air Quality Criteria for Particulate Matter: Second External Review Draft March 2001." *US Environmental Protection Agency, Office of Research and Development*.

Everson, P. and Morris, C. (2000). “Inference for multivariate Normal hierarchical models.” *Journal of the Royal Statistical Society, series B*, 62, 399–412.

Goldberg, M., Burnett, R., Valois, M., Flegel, K., and Bailar, J. (2003). “Associations between ambient air pollution and daily mortality among persons with congestive heart failure.” *Environ Research*, 91, 8–20.

Green, P., Jennison, C., and Seheult, A. (1985). “Analysis of Field Experiments by Least Square Smoothing.” *Journal of the Royal Statistical Society*, 47, 2, 299–315.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall, London UK.

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, 78S–90S.

Hastie, T., Tibshirani, R., and Buja, A. (1993). “Flexible Discriminant Analysis by Optimal Scoring.” Technical memorandum, ATT Bell Laboratories.

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

Katsouyanni, K., Schwartz, J., Spix, C., Touloumi, G., Zmirou, D., Zanobetti, A., and Wojtyniak, B. (1996). “Short term effects of air pollution on health: a European approach using epidemiologic time series data: the APHEA protocol.” *J Epidemiol Community Health*, 50, Supp 1: S12–8.

Katsouyanni, K., Touloumi, G., Samoli, E., Gryparis, A., LeTertre, A., Monopolis, Y., Rossi, G., Zmirou, D., Ballester, F., Boumghar, A., and Anderson, H. R. (2001). “Confounding and effect modification in the short-term effects of ambient particles on total mortality: Results from 29 European cities within the APHEA2 project.” *Epidemiology*, in press.

Katsouyanni, K., Touloumi, 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.

Klein, M., Flanders, W., and Tolbert, P. (2002). "Variances may be underestimated using available software for generalized additive models (abstract)." *American Journal of Epidemiology (supplement)*, 155, S106.

Lee, J., Kim, H., Song, H., Hong, Y., Cho, Y., Shin, S., Hyun, Y. J., and Kim, Y. (2002). "Air pollution and asthma among children in Seoul, Korea." *Epidemiology*, 13, 481–484.

Lumley, T. and Sheppard, L. (2003). "Time Series Analyses of Air Pollution and Health: Straining at Gnats and Swallowing Camels?" *Epidemiology*, 14, 13–14.

Marx, B. D. and Eilers, P. H. C. (1998). "Direct Generalized Additive Modeling With Penalized Likelihood." *Computational Statistics and Data Analysis*, 28, 193–209.

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

National Research Council (1998). "Research Priorities for Airborne Particulate Matter." *National Academy Press. Washington, DC.*

— (1999). "Research Priorities for Airborne Particulate Matter. Part II. Evaluating Research Progress and Updating the Portfolio." *National Academy Press. Washington, DC.*

— (2001). “Research Priorities for Airborne Particulate Matter. Part III. Early Research Progress.” *National Academy Press. Washington, DC*.

Nelder, J. A. and Wedderburn, R. W. M. (1972). “Generalized Linear Models.” *Journal of the Royal Statistical Society, Series A*, 135, 370–384.

Ramsay, T., Burnett, R., and Krewski, D. (2003). “The effect of concurvity in generalized additive models linking mortality and ambient air pollution.” *Epidemiology*, 14, 18–23.

Samet, J. (2000). “Epidemiology and Policy: The Pump Handle Meets the New Millenium.” *Epidemiologic Review*, 22, 145–154.

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, 11–12.

Samet, J., Zeger, S., Kelsall, J., Xu, J., and Kalkestein, L. (1998). “Does Weather Confound or Modify the Association of Particulate Air Pollution with Mortality ?” *Environmental Research*, 77, 9–19.

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.” *New England Journal of Medicine (with discussion)*, 343, 24, 1742–1757.

Samet, J. M., Zeger, S. L., and Berhane, K. (1995). *The Association of Mortality and Particulate Air Pollution*. Health Effects Institute, Cambridge, MA.

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 Part II: 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 Part I: Methods and Methodological Issues*. Health Effects Institute, Cambridge, MA.

Samet, J. M., Zeger, S. L., Kelsall, J., Xu, J., and Kalkstein, L. (1997). *Air pollution, weather and mortality in Philadelphia, In Particulate Air Pollution and Daily Mortality: Analyses of the Effects of Weather and Multiple Air Pollutants. The Phase IB report of the Particle Epidemiology Evaluation Project*. Health Effects Institute, Cambridge, MA.

Schwartz, J. (2000). "Assessing Confounding, Effect Modification, and Thresholds in the Associations between Ambient Particles and Daily Deaths." *Environmental Health Perspective*, 108, 563–568.

Schwartz, J., Spix, C., Touloumi, G., Bacharova, L., and Barumamdzadeh, T. e. a. (1996). "Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions." *J Epidemiol Community Health*, 50, Supp 1: S1–11.

Schwartz, J., Zanobetti, A., and Bateson, T. (2003). *A Special Report to the Health Effects Institute on the Revised Analyses of the NMMAPS II Data: Morbidity and Mortality among Elderly Residents of Cities with Daily PM Measurements*. The Health Effects Institute, Cambridge, MA.

Speckman, P. (1988). "Kernel Smoothing in Partial Linear Models." *Journal of the Royal Statistical Society Series B*, 50, 413–436.

Stieb, D., Judek, S., and Burnett, R. (2002). "Meta-analysis of time-series studies of air pollution and mortality: effects of gases and particles and the influence of cause of death, age, and season." *J Air Waste Manag Assoc.*, 52, 470–484.

The HEI Review Panels (2003). "Commentary to the HEI Special Report on the Revised Analyses of Time-Series Studies of Air Pollution and Health." *The Health Effects Institute, Cambridge, MA*.

Touloumi, G., Katsouyanni, K., Zmirou, D., and Schwartz, J. (1997). "Short-Term Effects of Ambient Oxidant Exposure on Mortality: A combined Analysis within the APHEA Project." *American Journal of Epidemiology*, 146, 177–183.

Zanobetti, A., Schwartz, J., and Dockery, D. (2000). "Airborne particles are a risk factor for hospital admissions for heart and lung disease." *Environmental Health Perspective*, 108, 1071–1077.

## Figure and Table Legends

Figure 1. Results of the simulation study when $g(t)$ is smoother than $f(t)$ (scenario A) and when $g(t)$ is more wiggly than $f(t)$ (scenario B), respectively. The first row shows the true $g(t)$ (solid line), the estimated $\widehat{g}(t)$ (dotted line), one realization of the pollution time series $x_t$. The second row shows the boxplots of the $N$ estimates ($\hat{\beta}_d^{\bullet,i} = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_d^{b,i}$) as function of $d$. The dots are plotted in correspondence of the unconditional average standard errors $\sqrt{\mathrm{UV}_d}$. The third row shows the unconditional squared bias ($\mathrm{USB}_d$) (triangles) and the unconditional variance ($\mathrm{UV}_d$) (dots) as function of $d$.

Figure 2. Four cities results: boxplots of $\widehat{\beta}^b(\alpha)$ for each city and for each value of $\alpha$. Solid and dotted horizontal lines are placed at $\widehat{\beta}_{\widehat{d}_1^\star,\widehat{d}_2^\star}$ and at 0, respectively.

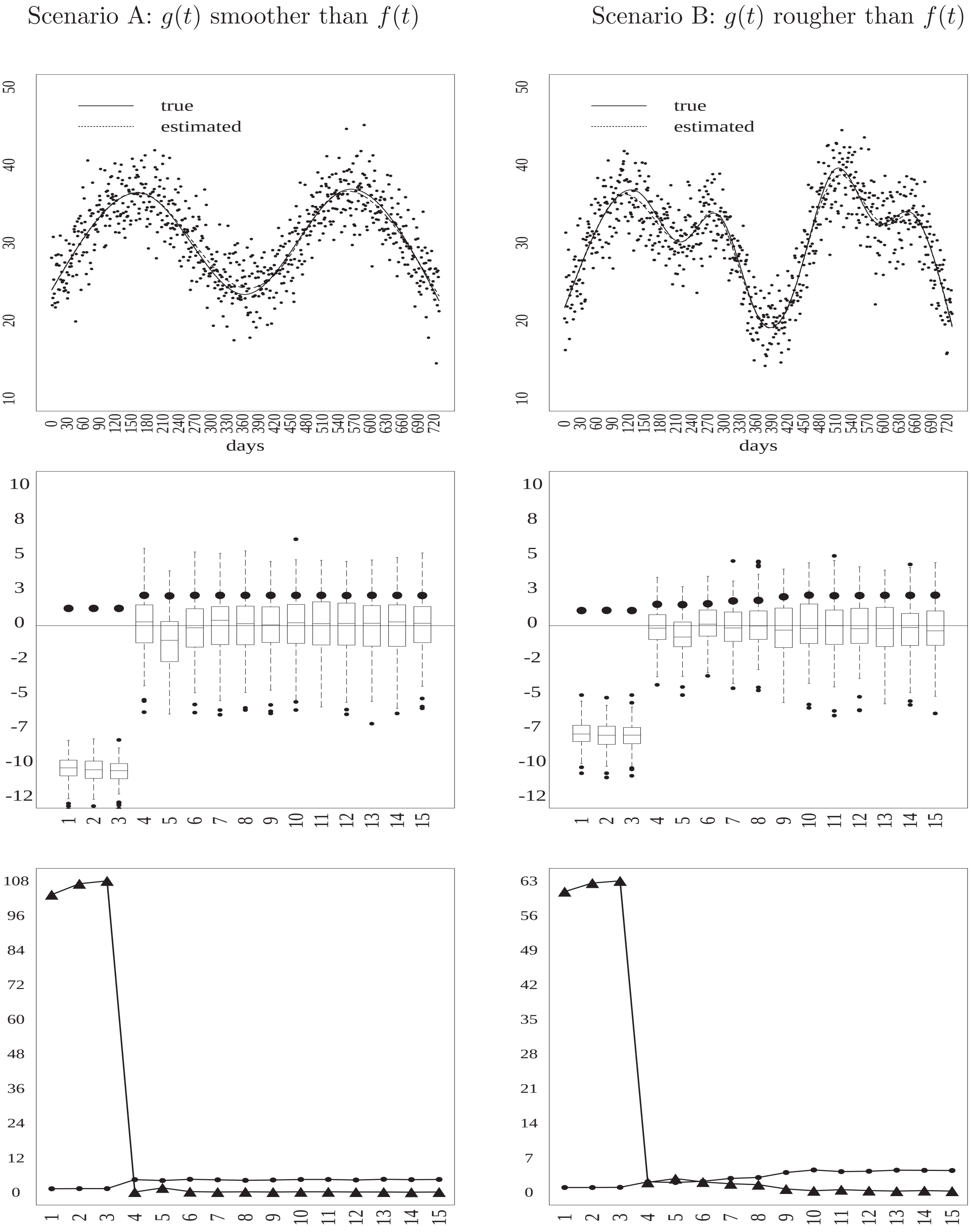
Figure 3. NMMAPS sensitivity analysis: top left panels show national average estimates (posterior means) as function of $\alpha$. Dots denote estimates under GAM with *approximated standard errors*, octagons denote estimates under GAM with *asymptotically exact standard errors*, and the triangles denote estimates under GLM. The grey polygon represents the 95% posterior intervals of the national average estimates under the GAM model with exact standard errors. The vertical segment is placed at $\alpha = 1$, e.g the degree of adjustment used in the NMMAPS model. The black curves in the top right panel denote the city-specific Bayesian estimates of the relative rates under GAM with asymptotically exact standard errors. Bottom panels show the posterior means of the average s.e. of $\widehat{\beta}^c$ ($\sqrt{\frac{1}{90}\sum_c v^c(\alpha)}$) (left) and posterior means of the heterogeneity parameters $\tau(\alpha)$ (right).

Table 1. Four cities results: $\widehat{d_1}, \widehat{d_2}$ denote the degrees of freedom that minimize GCV in the model that best predict $PM_{10}$ as smooth functions of time and temperature; $\widehat{\beta}_{\widehat{d}_1^\star,\widehat{d}_2^\star}$ denotes the estimate of the relative rate where the smooth functions of time and temperature are modelled

with $\widehat{d}_1^\star$ and $\widehat{d}_2^\star$, where $(\widehat{d}_1^\star, \widehat{d}_2^\star) = K \times (\widehat{d}_1, \widehat{d}_2)$, and $K = 3$ in Pittsburgh, Minneapolis, and Chicago, and $K = 2$ in Seattle.

| City | $\widehat{d}_1, \widehat{d}_2$ | $\widehat{\beta}_{\widehat{d}_1,\widehat{d}_2}$ | $\widehat{\beta}_{\widehat{d_1^\star},\widehat{d_2^\star}}$ |
|---|---|---|---|
| Pittsburgh | (30,6) | $0.27_{(-0.04,0.59)}$ | $0.24_{(-0.08,0.56)}$ |
| Minneapolis | (51,4) | $0.02_{(-0.52,0.57)}$ | $0.00_{(-0.57,0.57)}$ |
| Chicago | (51,6) | $0.29_{(0.06,0.53)}$ | $0.21_{(-0.03,0.44)}$ |
| Seattle | (140,10) | $0.16_{(-0.58,0.89)}$ | $-0.09_{(-0.90,0.71)}$ |

Table 1:


Scenario A: $g(t)$ smoother than $f(t)$

Scenario B: $g(t)$ rougher than $f(t)$

Figure 1:

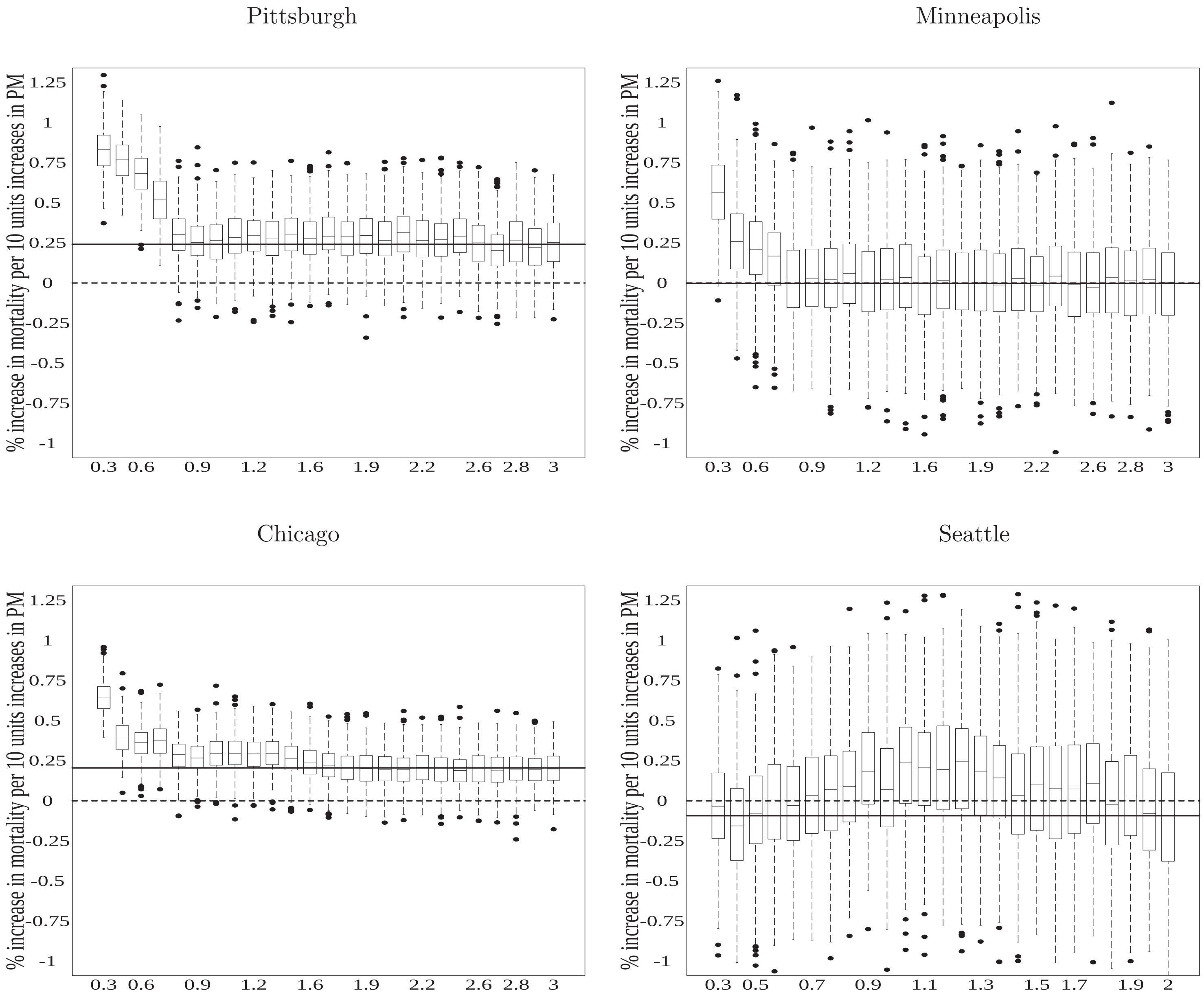
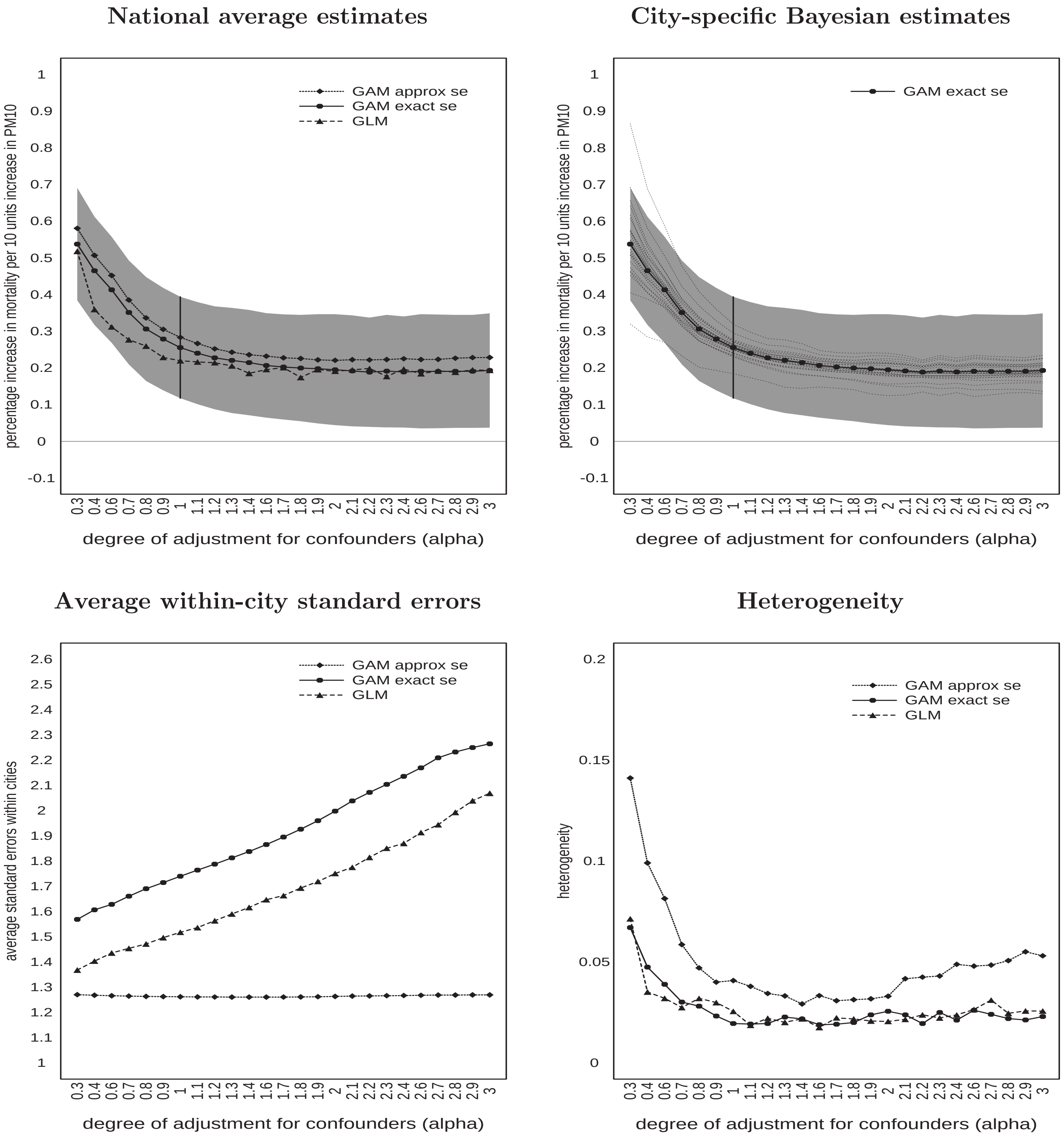

Pittsburgh
Minneapolis
Chicago
Seattle
% increase in mortality per 10 units increases in PM
1.25
1
0.75
0.5
0.25
0
-0.25
-0.5
-0.75
-1
0.3 0.6 0.9 1.2 1.6 1.9 2.2 2.6 2.8 3
0.3 0.5 0.7 0.9 1.1 1.3 1.5 1.7 1.9 2

Figure 2:



National average estimates

City-specific Bayesian estimates

Average within-city standard errors

Heterogeneity

Figure 3: