J. R. Statist. Soc. A (2006)
169, *Part* 2, *pp.* 179–203

Model choice in time series studies of air pollution and mortality

Roger D. Peng, Francesca Dominici and Thomas A. Louis

Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

[Received September 2004. Final revision July 2005]

Summary. Multicity time series studies of particulate matter and mortality and morbidity have provided evidence that daily variation in air pollution levels is associated with daily variation in mortality counts. These findings served as key epidemiological evidence for the recent review of the US national ambient air quality standards for particulate matter. As a result, methodological issues concerning time series analysis of the relationship between air pollution and health have attracted the attention of the scientific community and critics have raised concerns about the adequacy of current model formulations. Time series data on pollution and mortality are generally analysed by using log-linear, Poisson regression models for overdispersed counts with the daily number of deaths as outcome, the (possibly lagged) daily level of pollution as a linear predictor and smooth functions of weather variables and calendar time used to adjust for time-varying confounders. Investigators around the world have used different approaches to adjust for confounding, making it difficult to compare results across studies. To date, the statistical properties of these different approaches have not been comprehensively compared. To address these issues, we quantify and characterize model uncertainty and model choice in adjusting for seasonal and long-term trends in time series models of air pollution and mortality. First, we conduct a simulation study to compare and describe the properties of statistical methods that are commonly used for confounding adjustment. We generate data under several confounding scenarios and systematically compare the performance of the various methods with respect to the mean-squared error of the estimated air pollution coefficient. We find that the bias in the estimates generally decreases with more aggressive smoothing and that model selection methods which optimize prediction may not be suitable for obtaining an estimate with small bias. Second, we apply and compare the modelling approaches with the National Morbidity, Mortality, and Air Pollution Study database which comprises daily time series of several pollutants, weather variables and mortality counts covering the period 1987–2000 for the largest 100 cities in the USA. When applying these approaches to adjusting for seasonal and long-term trends we find that the Study's estimates for the national average effect of PM_{10} at lag 1 on mortality vary over approximately a twofold range, with 95% posterior intervals always excluding zero risk.

Keywords: Air pollution; Log-linear regression; Mortality; Semiparametric regression; Time series

1. Introduction

Numerous time series studies have indicated a positive association between short-term variation in particulate matter (PM) and daily mortality counts (see for example Pope *et al.* (1995), Dockery and Pope (1996), Goldberg *et al.* (2003), Bell *et al.* (2004) and references therein). Multicity studies such as the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) (Samet *et al.*, 2000a, b), the 'Air pollution and health: a European approach' study (Katsouyanni *et al.*, 2001; Samoli *et al.*, 2002) and analyses of Canadian cities (Burnett *et al.*, 1998; Burnett and Goldberg, 2003) have added to the mounting evidence of the adverse health effects of fine

Address for correspondence: Roger D. Peng, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205, USA.
E-mail: rpeng@jhsph.edu

 0964–1998/06/169179

particles, even at levels that are below current regulatory limits. In the USA, these studies have played an important role in setting standards for acceptable levels of ambient PM. In particular, the NMMAPS played a central role in the Environmental Protection Agency's development of national ambient air quality standards for the six 'criteria' pollutants defined by the Environmental Protection Agency (Environmental Protection Agency, 1996, 2003).

The critical role of the NMMAPS in the development of the air quality standards attracted intense scrutiny from the scientific community and industry groups regarding the statistical models that are used and the methods that are employed for adjusting for potential confounding. Confounding occurs when an attribute that is associated with an outcome is also associated with the exposure of interest but is not a result of the exposure. In time series studies, we are primarily concerned with potential confounding by factors that vary on similar timescales as pollution or mortality. Although collectively strengthening the epidemiologic evidence of the adverse health effects of PM, the proliferation of time series studies employing different approaches to modelling and adjusting for confounding highlighted the critical need to assess the statistical properties of these approaches.

The different sources of potential confounding in time series studies of air pollution and mortality can be broadly classified as either *measured* or *unmeasured*. Important measured confounders include weather variables such as temperature and dewpoint temperature. Daily temperature measurements are readily available for metropolitan areas in the USA and numerous studies have demonstrated a relationship between temperature and mortality which is generally positive for warm summer days and negative for cold winter days (e.g. Curriero *et al.* (2002)). One approach to adjusting for confounding by temperature is to include non-linear functions of current and previous day temperature (and dewpoint) in the model (Schwartz, 1994a; Kelsall *et al.*, 1997; Samet *et al.*, 1998). Welty and Zeger (2005) developed a rich class of distributed lag models that were specifically targeted at adjusting for temperature in multicity time series studies of air pollution and mortality. This class of models includes a variety of predictors such as running means of temperature, non-linear functions of running means, multiple lags of temperature and interactions between temperature at different lags. They applied their models to the NMMAPS database and found that the national average estimate of the effect of PM_{10} (PM with aerodynamic diameter less than 10 μm) on total non-accidental mortality is robust to a large class of statistical models that are used to adjust for potential confounding by temperature and dewpoint temperature. Building on these findings, in this paper we focus on the problem of controlling for unmeasured confounders, i.e. seasonal and long-term trends.

Unmeasured confounders are factors that influence mortality and vary with time in a manner that is similar to air pollution. These factors produce seasonal and long-term trends in mortality that can confound the relationship between mortality and air pollution. Influenza and respiratory infections might reasonably be considered among the most important, usually unmeasured or not readily available confounders which produce seasonal patterns in mortality. Typically, epidemic respiratory infections occur from late autumn to early spring and influenza epidemics occur in the same interval but with highly variable timing. The net effect of a respiratory virus is to increase mortality overall, explaining much of the higher mortality in winter months. Since air pollution levels also have a strong seasonal pattern, such respiratory virus epidemics are likely to confound the relationship between air pollution and mortality. Daily time series of mortality counts can also be affected by population level trends in survival (including increased or decreased access to improved medical care), changes in population size and trends in the occurrence of major diseases. These long-term trends could coincide with recent declines in a number of pollution indicators (e.g. total suspended particles and then PM_{10}).

A common approach to adjusting for seasonal and long-term trends is to use semiparametric models which incorporate a smooth function of time. The use of nonparametric smoothing in time series models of air pollution and health was suggested in Schwartz (1994a), where generalized additive Poisson models were used with LOESS smooths of time, temperature, dewpoint temperature and PM_{10}. This approach can be thought of as regressing residuals from the smoothed dependent variable on residuals from the smoothed regressors. In this setting, the smooth function of time serves as a linear filter on the mortality and pollution series and removes any seasonal or long-term trends in the data. Several alternatives for representing the smooth functions have been applied including smoothing splines, penalized splines and parametric (natural) splines (Dominici *et al*., 2002; Ramsay *et al*., 2003; Schwartz *et al.*, 2003; Touloumi *et al*., 2004; Health Effects Institute, 2003). The smooth function of time naturally accounts only for potential confounding by factors which vary smoothly with time. Factors which vary on shorter timescales may also confound the relationship between air pollution and mortality and controlling for them is an important concern.

The inclusion of a smooth function of time in a regression model introduces important statistical issues. We generally do not know precisely the complexity of the seasonal and long-term trends in the mortality time series or in the pollution time series. Therefore, a controversial issue is determining how much smoothness we should allow for the smooth function of time. This decision is critical because it determines the amount of residual temporal variation in mortality that is available to estimate the air pollution effect. Oversmoothing the series (thereby undersmoothing the residuals) can leave temporal cycles in the residuals that can produce confounding bias; undersmoothing the series (thereby oversmoothing the residuals) can remove too much temporal variability and potentially attenuate a true pollution effect. Current approaches to choosing the amount of smoothness include automatic, data-driven methods which choose the degree of smoothness by minimizing a goodness-of-fit criterion and methods based on prior knowledge of the timescales where confounding is more likely to occur.

In this paper we provide a comprehensive characterization of model choice and model uncertainty in time series studies of air pollution and mortality, focusing on confounding adjustment for seasonal and long-term trends. We first identify analytical approaches that are used commonly in air pollution epidemiology for modelling the smooth function of time and for selecting its degrees of freedom. We then introduce a statistical framework that allows us to compare and evaluate critically the statistical properties of each modelling approach by illustrating its theoretical properties and by simulation studies. Finally, we apply the different approaches for confounding adjustment to the NMMAPS database containing daily mortality, pollution and weather data for 100 US cities covering the period 1987–2000. Here, we quantify model uncertainty in the most recent national average estimates of the short-term effects of PM on mortality.

2. Methods and model choice

Given time series data on pollution levels, mortality counts and other variables, we make use of the statistical model

$$\begin{aligned} Y_t &\sim \text{Poisson}(\mu_t), \\ \log(\mu_t) &= \beta_0 + \beta x_t + f(t) + q(z_t) + w_t. \end{aligned} \tag{1}$$

Y_t is the mortality count for day t; f is a smooth function of the time variable t; z_t represents an observed time-varying variable such as temperature and q is a (smooth) function of that variable; w_t is some other linear term such as a day of the week or holiday indicator. Our goal

is to estimate the parameter β, the association between air pollution x_t and mortality Y_t, in the presence of unobserved, time-varying confounding factors. We assume that these factors potentially influence μ_t ($\mathbb{E}(Y_t)$) via the smooth function f and, to produce confounding, are associated with x_t through another smooth function g, via

$$x_t = g(t) + \xi_t, \tag{2}$$

where $\xi_t \sim \mathcal{N}(0, \sigma^2)$ and $\sigma^2 > 0$. If f and g are correlated at similar timescales, confounding bias can occur because mortality and pollution vary with time in a similar manner. Correlation between f and g in a nonparametric setting is sometimes referred to as *concurvity*, essentially collinearity between non-linear transformations of predictors, and is the nonparametric analogue of collinearity in standard multiple-regression analysis (Buja *et al.*, 1989; Donnell *et al.*, 1994). The strength of the concurvity between f and g is determined by the parameter σ^2, which we assume is strictly greater than 0. If $\sigma^2 = 0$, then f and g are perfectly correlated and the problem of estimating β is not identifiable. Our statistical and epidemiological target is to determine the degree of smoothness of f that maximally reduces the confounding bias in $\hat{\beta}$, the estimate of the pollution coefficient β, for $\sigma^2 > 0$.

With a model set-up such as expression (1), to estimate β, we must choose how to represent the smooth function f and then decide on the amount of smoothness that is allowed for f. In practice f is typically represented by a series of basis functions and the smoothness is controlled by the number of basis functions or, more generally, a notion of 'degrees of freedom'.

2.1. Representing f

Common choices for representing the smooth function f in model (1) include natural splines, penalized splines and smoothing splines. (Other less common choices are LOESS smoothers or harmonic functions.) The first is fully parametric, whereas the last two may be considered more flexible. With natural splines, we construct a spline basis with knots at fixed locations throughout the range of the data and the choice of knot locations can have a substantial effect on the resulting smooth. Smoothing splines and penalized splines circumvent the problem of choosing the knot locations by constructing a very large spline basis and then penalizing the spline coefficients to reduce the effective number of degrees of freedom. Smoothing splines place knots at every (unique) data point and are sometimes referred to as full rank smoothers because the size of the spline basis is equal to the number of observations. Penalized splines, sometimes called low rank smoothers, are more general in their definition in that both the size of the spline basis and the location of the knots can be specified. Low rank smoothers can often afford significant computational benefits when applied to larger data sets such as those used here. Appendix A provides an overview of the different methods that are used here; a comprehensive treatment can be found in Ruppert *et al.* (2003).

We employ three commonly used software implementations to fit models by using the different spline bases.

(a) GLM-NS: the `glm` function in R (R Development Core Team, 2003) is used with natural cubic splines to represent f. The number of degrees of freedom for the spline basis is specified via the `df` argument of the `ns` function (in the `splines` package).
(b) GAM-R: the `gam` function in R (from the `mgcv` package) is used with penalized cubic regression splines to represent f. This function allows the user to specify the dimension of the basis (before penalization) as well as a penalty parameter. In our simulations and data analysis we use a basis dimension that is equal to 40 times the number of years of data. The number 40 per year of data was chosen because it was considered far more

degrees of freedom than would be necessary to remove seasonal and long-term variation in pollution and, hence, some penalization would be required. The implementation of `gam` in R uses a form of iteratively reweighted least squares to fit the model and standard errors for the regression coefficients can be obtained in a straightforward manner. The methods and software are described in Wood (2000, 2001).

(c) GAM-S: the `gam` function in S-PLUS is used with smoothing splines to represent f. This function is *not* the same as the `gam` function in R. Here, the user specifies the target number of degrees of freedom that are desired. The size of the basis does not need to be specified since it is determined by the number of unique data points. The S-PLUS implementation of `gam` uses backfitting to estimate the smooth terms and we use the strict convergence criteria that were suggested in Dominici *et al.* (2002). Standard errors are obtained by using the `gam.exact` software of Dominici *et al.* (2004).

Because of the close relationship between penalized splines and smoothing splines (see Appendix A) we compare only the GLM-NS and GAM-R methods in the simulation study. Furthermore, preliminary comparisons of the penalized spline and smoothing spline methods indicated that they performed similarly. For the analysis of the NMMAPS data in Section 4 we compare all three methods.

2.2. Selecting the degrees of freedom for f

Given a particular representation of f described in Section 2.1, we must then choose the amount of smoothness to allow for f. We examine model selection approaches that have already been used extensively by investigators in the area of time series modelling of air pollution and health data. A general strategy is to use a data-driven method and to select the number of degrees of freedom (df) which optimizes a particular criterion. For example, one approach is to choose the df which leads to optimal prediction of the mortality outcome series (e.g. Burnett and Goldberg (2003)) and another is to select the df which best predicts the pollution series (Dominici *et al.*, 2004). A third strategy is to minimize the autocorrelation in the residuals (e.g. Schwartz (2000), Katsouyanni *et al.* (2001), Samoli *et al.* (2002, 2003) and Touloumi *et al.* (2004)). With each of these approaches, a number of Poisson regression models are fitted using a range of df-values (other covariates such as weather variables and the pollutant variable are included). Then, for each fitted model, a model selection criterion is evaluated with the 'optimal' df being that which minimizes the criterion. In multicity studies, this approach can lead to a different df selected for each city (using a common criterion across cities), potentially allowing city-specific characteristics of the data to influence the estimated smoothness of f.

Another approach which we examine here is to use a fixed degrees of freedom, perhaps based on biological knowledge or previous work. For multicity studies, this approach generally leads to fitting the same model to data from each city. The original NMMAPS analyses took this approach and used 7 degrees of freedom per year of data (Samet *et al.*, 2000a). One can explore the sensitivity of $\hat{\beta}$ by varying the df that is used in the model(s) and examining the associated changes in $\hat{\beta}$.

In summary, we explore the following strategies for deciding on an appropriate number of degrees of freedom (df) for f.

(a) *Fixed degrees of freedom*: choose a fixed df based on biological knowledge or previous work and include a sensitivity analysis to explore the variability of $\hat{\beta}$ with respect to df. For the sensitivity analysis we estimate β for $\mathrm{df} = 1, 2, \ldots, 20$ per year of data.

(b) *Akaike information criterion*: choose the df that minimizes the Akaike information criterion AIC (Akaike, 1973). AIC is commonly used for selecting particular covariates and

has been applied to the smooth function of time. For a model with df degrees of freedom, AIC is defined as

$$\mathrm{AIC}(\mathrm{df}) = -2(\text{maximum log-likelihood}) + 2\,\mathrm{df}.$$

(c) *Bayesian information criterion*: choose the df that minimizes the criterion of Schwarz (1978). This criterion is often referred to as the Bayesian information criterion BIC and is sometimes used as an approximation to the posterior model weight, for example, in Bayesian model averaging (e.g. Daniels *et al*. (2000) and Clyde (2000)). BIC can be written as

$$\mathrm{BIC}(\mathrm{df}) = -2(\text{maximum log-likelihood}) + \log(n)\,\mathrm{df}$$

where n is the number of observations.

(d) *Minimum residual autocorrelation*: choose the df that minimizes the autocorrelation in the residuals. In practice we can minimize the sum of the absolute value of the partial autocorrelation function (PACF) of the residuals for a fixed number of lags. An alternative is choosing df by using a test for white noise in the residuals (e.g. Goldberg *et al*. (2001) and others). Although this approach is used in the literature, we do not explore it here because common tests for white noise (such as the portmanteau test) are either functions of the autocorrelation function coefficients or are closely related (Brockwell and Davis, 2002). Hence, the df which minimizes the sum of the absolute value of the PACF coefficients should correspond closely to the df that leads a test for white noise to fail to reject the null hypothesis.

(e) *GCV-PM*$_{10}$: choose the df that best predicts the *pollution* series, as measured by generalized cross-validation (Gu, 2002). This approach is a simplified version of the mean-squared error minimization procedure that was described in Dominici *et al*. (2004).

3. Simulation study

In this section we describe a simulation study that was designed to assess the bias and mean-squared error of $\hat{\beta}$ under different basis representations for f and the five approaches to selecting df that were described in Section 2. Our goal is to generate data from confounding scenarios that are comparable with situations found in real data and to evaluate the estimation procedures in each of these scenarios. The definition of the scenarios relies on the timescales at which confounding occurs and the strength of the concurvity between the pollutant series and the seasonal trend. All the simulations were conducted in R by using the `glm` and `ns` functions to fit natural spline models and the `gam` function in the `mgcv` package to fit penalized spline models.

Our statistical framework for the simulations is

$$\left.\begin{aligned} Y_t &\sim \mathrm{Poisson}(\mu_t), \\ \log(\mu_t) &= \beta_0 + \beta\,\mathrm{PM}_t + f(t) + q(\mathrm{temp}_t), \\ \mathrm{PM}_t &= g(t) + r(\mathrm{temp}_t) + \xi_t \qquad \xi_t \sim \mathcal{N}(0, \sigma^2), \end{aligned}\right\} \tag{3}$$

where PM_t and temp_t are the PM_{10} and temperature time series respectively. We assume that f and g have the natural spline representations

$$\begin{aligned} f(t) &= \sum_{j=1}^{m_1} a_j B_j(t), \\ g(t) &= \sum_{j=1}^{m_2} b_j H_j(t), \end{aligned} \tag{4}$$

where the B_j and H_j are known basis functions and m_1 and m_2 are the degrees of freedom for f and g respectively. The functions q and r also have natural spline representations with n_1 and n_2 degrees of freedom.

To simulate mortality and pollution data, we first specify values m_1, m_2, n_1 and n_2. Then, we fit a log-linear Poisson regression model to the Minneapolis–St Paul total non-accidental mortality data to obtain estimates of the spline coefficients $a_1, \ldots, a_{m_1}$ and a standard linear regression model to the PM_{10}-data to obtain estimates of the spline coefficients $b_1, \ldots, b_{m_2}$. Data from Minneapolis–St Paul for the years 1987–1994 were used because the city has daily measurements of PM_{10} and a sufficient number of deaths to produce a stable estimated effect of PM_{10} on mortality. We also estimate the residual variance from the PM_{10} regression model for the Minneapolis–St Paul data and call it σ_0^2. The parameter σ_0^2 is used later to control how much concurvity will exist in the simulated data.

All the parameters that are estimated from the Minneapolis–St Paul data are then treated as the 'true' coefficients from which to simulate. The framework in model (3) ensures that some concurvity will exist between the simulated mortality and pollution data, the strength of which we can control via the specification of σ^2, the variance of ξ_t in expression (3). For example, if we set $\sigma^2 = \sigma_0^2/10$, this would produce simulated data with high concurvity. Note that we do not generate temperature data; they remain fixed in each of the simulations.

We simulate the following four confounding scenarios.

(a) $g(t)$ is smoother than $f(t)$; moderate concurvity. Confounding bias might occur because longer cycles in the air pollution are correlated with the longer cycles in mortality and the amount of correlation depends on the variance σ^2. However, the mortality counts might also be affected by factors that vary at shorter cycles than pollution. Here we set $m_1 = 7 \times 8 = 56$, $m_2 = 4 \times 8 = 32$, $n_1 = 6$, $n_2 = 3$ and $\sigma^2 = \sigma_0^2$.

(b) $g(t)$ is smoother than $f(t)$; high concurvity. This is the same as in scenario (a) except that we set $\sigma^2 = \sigma_0^2/10$. Here the pollution variable PM_t is very tightly correlated with the smooth function of time f.

(c) $g(t)$ is rougher than $f(t)$; moderate concurvity. Confounding bias might occur because longer cycles in air pollution are correlated with the longer cycles in the mortality counts. Temporal variation in pollution levels might also be affected by factors that vary at shorter cycles than the mortality counts. Here we set $m_1 = 32$, $m_2 = 56$, $n_1 = 3$, $n_2 = 6$ and $\sigma^2 = \sigma_0^2$.

(d) $g(t)$ is rougher than $f(t)$; high concurvity. This is the same as in scenario (c) except that we set $\sigma^2 = \sigma_0^2/10$.

The four simulation scenarios are summarized in Table 1. Our simulation framework does not address the issue of measurement error in the pollutant variable. Since such error can in some

Table 1. Simulation scenarios†

Scenario	*Concurvity*	σ^2	m_1 *(df for f)*	m_2 *(df for g)*
$g(t)$ smoother than $f(t)$	Moderate	σ_0^2	56	32
$g(t)$ smoother than $f(t)$	High	$\sigma_0^2/10$	56	32
$g(t)$ rougher than $f(t)$	Moderate	σ_0^2	32	56
$g(t)$ rougher than $f(t)$	High	$\sigma_0^2/10$	32	56

†$\sigma_0^2 = 186.7$ for scenarios where $g(t)$ is smoother than $f(t)$ and $\sigma_0^2 = 182.2$ for scenarios where $g(t)$ is rougher than $f(t)$.

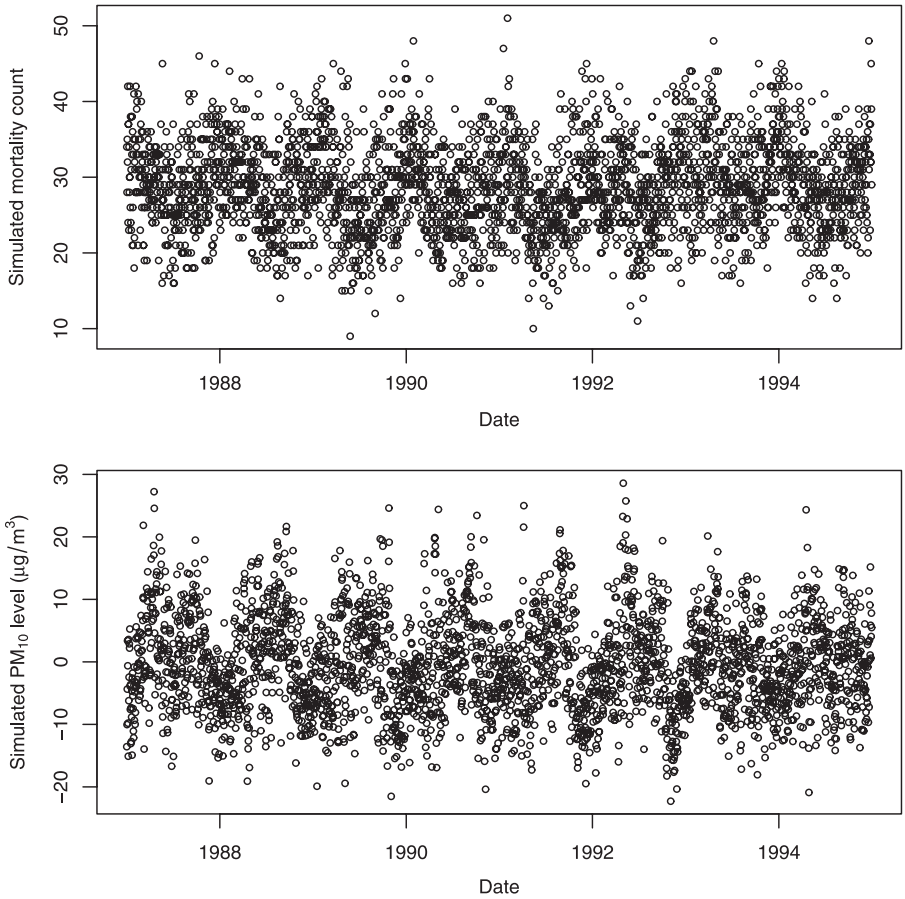

Fig. 1. Example of simulated mortality and PM_{10}-data: the negative values in the PM_{10}-series come from the original data being represented as deviations from an overall mean; in this example, g is smoother than f and there is high concurvity

situations attenuate the estimated pollution effect, it may be useful in the future to employ a more elaborate simulation framework to investigate in detail the effect of measurement error.

We generate mortality and pollution data from these scenarios assuming no pollution effect ($\beta = 0$). For each scenario that is listed in Table 1 we simulate $N = 500$ data sets and fit a Poisson regression model to each by using either natural splines or penalized splines for a range of values of df. That range was df = 1–20 per year of data in the simulated data set, which in this case was 8 years. Fig. 1 shows one of the simulated data sets for the scenario where g is smoother than f and there is high concurvity. To each simulated data set we apply the five df selection methods that were described in Section 2 and investigate under which circumstances we would wrongly report a statistically significant air pollution effect.

Fig. 2 shows box plots of the 500 estimates of β obtained by using df = 1–20 per year in the smooth function of time. Figs 2(a)–2(d) show estimates that were obtained by using natural splines and Figs 2(e)–2(h) show the results of using penalized splines to represent f. Figs 2(c), 2(d), 2(g) and 2(h) show the estimates that were obtained under the high concurvity scenario. In general, although the variance of the estimates tends to increase as the number of degrees of freedom for f is increased, the decrease in bias is far more dramatic. Under moderate concurvity

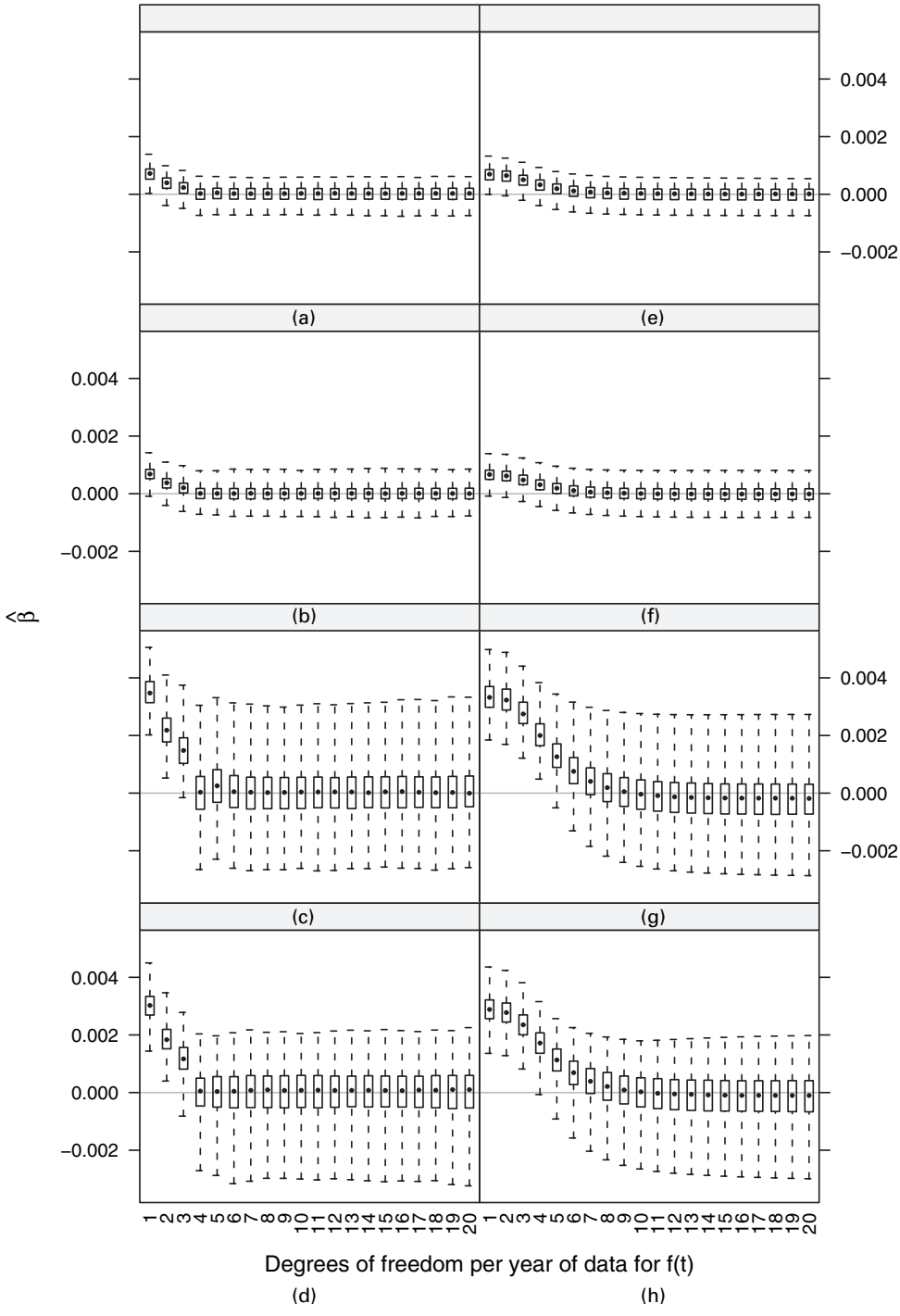

Fig. 2. Sensitivity analysis of $\hat{\beta}$ (distribution of $\hat{\beta}$ over 250 simulations using df = 1–20 per year in the smooth function of time f (the true $\beta = 0$)): (a) GLM-NS (*g* smoother); (b) GLM-NS (*g* rougher); (c) GLM-NS (*g* smoother; high concurvity); (d) GLM-NS (*g* rougher; high concurvity); (e) GAM-R (*g* smoother); (f) GAM-R (*g* rougher); (g) GAM-R (*g* smoother; high concurvity); (h) GAM-R (*g* rougher; high concurvity)

(Figs 2(a), 2(b), 2(e) and 2(f)) the bias in the estimates is only serious for df between 1 and 4 for natural splines (between 1 and 6 for penalized splines).

The apparent decrease in the bias of $\hat{\beta}$ with increasing df is explained in Dominici *et al.* (2004) for the natural spline case and explored by Rice (1986) and Speckman (1988) in the nonparametric setting. Dominici *et al.* (2004) showed that for natural splines, if we select df to be equal to the df that is necessary to represent the g-function in model (3), then $\hat{\beta}$ is either unbiased (when g is rougher than f) or asymptotically unbiased (when g is smoother than f). For example, with g rougher than f, we should see very little bias in $\hat{\beta}$ for df $\geqslant 7$ per year. In the nonparametric setting, Rice and Speckman both showed that, to obtain an estimate of β whose bias converges at the usual parametric rate, we must undersmooth the estimate of f (see Appendix A.1 for more details). An important conclusion here is that, when using either natural splines or penalized splines, the amount of smoothing in f that is required to obtain an estimate of β with small bias could be less than the amount of smoothing that is required to obtain a good estimate of f alone (see also Green and Silverman (1994), chapter 4).

Under high concurvity, the differences between using natural splines and penalized splines are greater. For natural splines, the bias drops rapidly between df $= 1$ and df $= 4$ per year and is stable afterwards. For penalized splines, the bias drops much more slowly and does not appear to level off until df $= 9$ or df $= 10$ per year. In general, the estimates of β appear to be less sensitive to the relationship between the g- and f-functions (i.e. g smoother or rougher) than to the amount of concurvity in the data or the basis representation that is used.

In our comparison of the model selection criteria that were described in Section 2.2, for each simulated data set and criterion, we obtain a 'best' df, call it $\widehat{\text{df}}$, that is the value of df associated with the fitted model which minimizes the criterion. The estimate of β that is chosen by the model selection criterion for data set i is $\hat{\beta}^{(i)}_{\widehat{\text{df}}}$. We can then estimate the bias, standard error and root-mean-squared error (RMSE) of $\hat{\beta}_{\widehat{\text{df}}}$ from the simulation output for a particular model selection criterion and choice of basis. Clearly, the RMSE for a criterion depends on an effective balance between the bias and variance of the estimates.

The average bias, standard error and RMSE (all multiplied by 1000) for $\hat{\beta}$ that were selected by each of the criteria–bases under the various scenarios are shown in Table 2. Along the rows labelled 'df $= m_1$', Table 2 also shows the same results for the estimates of β when the df that is used to generate the data (whose specific values are shown in Table 1) is used as the 'best' df rather than minimizing one of the model selection criteria. Under moderate concurvity, each of the four data-driven methods performs reasonably well with respect to the RMSE with BIC always having the largest RMSE. As expected, all the methods perform worse under high concurvity, with BIC having an RMSE that is more than twice as large as the other methods in some instances.

Table 2 also shows the contribution of bias and variance to the RMSEs of the estimates of β that were obtained via the model selection criteria. Generally, estimates from all the criteria incur more bias when using penalized splines for the smooth function of time as opposed to natural splines. GCV-PM_{10} is very nearly unbiased in all the scenarios. The largest bias (0.159) occurs with penalized splines, under high concurvity and when g is smoother than f. AIC has a relatively small bias under the moderate concurvity scenarios but tends to incur more bias than GCV-PM_{10} under the high concurvity scenarios (particularly when penalized splines are used). The price for using GCV-PM_{10} over the other methods appears to be an increase in the standard error of the estimates in some cases.

The PACF criterion performs reasonably well under moderate concurvity but has a large bias under high concurvity, particularly when using penalized splines. However, the relative increase in bias for the PACF criterion when going from the moderate concurvity to the high concurvity

Table 2. Average bias, standard error and RMSE of $\hat{\beta}$ (all times 1000) from 250 simulations†

	Results for natural splines				*Results for penalized splines*			
	Moderate concurvity		*High concurvity*		*Moderate concurvity*		*High concurvity*	
	g(t) smoother	*g(t) rougher*	*g(t) smoother*	*g(t) rougher*	*g(t) smoother*	*g(t) rougher*	*g(t) smoother*	*g(t) rougher*
Bias (×1000)								
AIC	0.012	0.012	0.026	0.119	0.061	0.152	0.421	1.000
PACF	0.059	0.305	0.401	1.701	0.383	0.570	2.359	2.675
BIC	0.492	0.471	3.302	2.782	0.663	0.652	3.343	2.884
GCV-PM_{10}	0.021	0.002	0.014	0.034	0.121	0.041	0.159	−0.030
$df = m_1$	0.013	0.005	0.018	0.024	0.059	0.306	0.416	1.715
Standard error (×1000)								
AIC	0.255	0.258	0.823	0.803	0.252	0.252	0.752	0.692
PACF	0.268	0.299	1.002	0.833	0.267	0.253	0.770	0.543
BIC	0.305	0.308	0.798	0.730	0.257	0.243	0.540	0.501
GCV-PM_{10}	0.255	0.258	0.818	0.805	0.249	0.253	0.741	0.742
$df = m_1$	0.256	0.253	0.819	0.695	0.250	0.243	0.712	0.541
RMSE (×1000)								
AIC	0.361	0.364	1.164	1.142	0.362	0.388	1.144	1.399
PACF	0.383	0.521	1.473	2.068	0.538	0.673	2.598	2.783
BIC	0.654	0.641	3.490	2.968	0.756	0.737	3.429	2.969
GCV-PM_{10}	0.361	0.365	1.157	1.138	0.372	0.361	1.060	1.049
$df = m_1$	0.361	0.357	1.158	0.982	0.359	0.460	1.089	1.878

†Each column represents a scenario that is determined by the basis that is used for fitting (natural splines or penalized splines), the concurvity in the simulated data and the relationship between $g(t)$ and $f(t)$, i.e. $g(t)$ smoother or rougher than $f(t)$.

scenarios is comparable with the other criteria. The BIC-criterion performs poorly under all the scenarios. The larger penalty that is associated with BIC generally leads to using few degrees of freedom which, from Fig. 2, can produce estimates with high bias.

4. National Morbidity, Morbidity, and Air Pollution Study data analysis

We apply our methods to the NMMAPS database which comprises daily time series of air pollution levels, weather variables and mortality counts. The original study examined data from 90 cities for the years 1987–1994 (Samet *et al.*, 2000a, b). The data have since been updated to include 10 more cities and six more years of data, extending the coverage until the year 2000. The entire database is available via the NMMAPSdata R package (Peng and Welty, 2004) which can be downloaded from the Internet-based health and air pollution surveillance system Web site at `http://www.ihapss.jhsph.edu/`.

The full model that is used in the analysis for this section is larger than the simpler model that was described in Section 3. We use an overdispersed Poisson model where, for a single city,

$$\log\{\mathbb{E}(Y_t)\} = \text{age-specific intercepts} + \text{day of week} + \beta\,\text{PM}_t + f(\text{time, df}) + s(\text{temp}_t, 6) + s(\text{temp}_{1-3}, 6) + s(\text{dewpoint}_t, 3) + s(\text{dewpoint}_{1-3}, 3).$$

Here, f is the smooth function of time represented with different bases and $s(\cdot, d)$ indicates a smooth function with d degrees of freedom. In addition to a smooth function of time and the PM_{10}-series, the model includes smooth functions of temperature, dewpoint temperature and three day running means of each (denoted by the subscripts 1–3). There is also an indicator variable for the day of the week and a separate intercept for each age category (less than 65, 65–74 and 75 years old or older).

For each city, we choose each of the three fitting procedures (i.e. representations of the smooth function of time) that were described in Section 2.1 and fit an overdispersed Poisson model. We then minimize one of the criteria described in Section 2.2 and obtain a best df, call it $\widehat{\mathrm{df}}$, with which we obtain an estimate $\hat{\beta}_{\widehat{\mathrm{df}}}$ for that city. This process is then repeated for all 100 cities in the database to obtain $\hat{\beta}^{(1)}_{\widehat{\mathrm{df}}}, \ldots, \hat{\beta}^{(100)}_{\widehat{\mathrm{df}}}$ and their standard errors. These city-specific estimates are pooled using a two-level hierarchical normal model (similar to that used in Dominici *et al.* (2000)) with flat priors on the overall estimate and the between-city covariance matrix (Everson and Morris, 2000a, b). The result is a 'national average estimate' summarizing the effect of PM_{10} on mortality for the 100 cities. We run this entire process for each of the three fitting procedures and three model selection criteria: AIC, PACF and GCV-PM_{10}. For the overdispersed Poisson models we use a modified AIC of the form (Hastie and Tibshirani, 1990)

$$\mathrm{AIC} = -2(\text{maximum log-likelihood}) + 2\,\mathrm{df}\,\hat{\phi},$$

where $\hat{\phi}$ is the estimated dispersion parameter.

Table 3 shows the results of applying the model selection criteria and using different representations of the smooth function of time for the NMMAPS data. The estimates that are presented are the national average estimates of the percentage increase in mortality for an increase in PM_{10} of 10 μg m^{-3} at lag 1. The results are consistent with what we observed in the simulation studies—AIC and GCV-PM_{10} produce very similar estimates whereas the PACF estimates are somewhat larger. The estimates that were obtained by AIC and GCV-PM_{10} are comparable with the estimates that were reported in previous NMMAPS analyses (e.g. Dominici *et al.* (2002, 2003) and Peng *et al.* (2005)), although with smaller 95% posterior intervals due to the additional data that are used in the current analysis.

A problem arises with the PACF procedure when cities with a regular pattern of missing PM_{10}-data are included (something which is common with US data). In particular, for cities where PM_{10} is measured only once every 6 days, we can only estimate the autocorrelation of the residuals at lag 6. The national average estimates in the third column of Table 3 were computed by ignoring the 1-in-6 pattern in the data. Cities with sporadic missing PM_{10}-values do not cause a problem in computing the PACF.

Fig. 3 shows a sensitivity analysis of the national average estimate with respect to the number of degrees of freedom per year assigned to the smooth function of time. In Fig. 3, rather than minimize one of the model selection criteria and obtain an optimal df in each city, we use a

Table 3. National average estimates and 95% posterior intervals of the percentage increase in mortality with an increase in PM_{10} of 10 μg m^{-3} at lag 1 by using different model selection criteria and representations of the smooth function of time, $f(t)$

Method	*AIC*	*PACF*	*GCV-PM*$_{10}$
GLM-NS (natural splines)	0.20 (0.11, 0.29)	0.25 (0.14, 0.36)	0.20 (0.10, 0.29)
GAM-R (penalized splines)	0.25 (0.16, 0.34)	0.35 (0.24, 0.46)	0.26 (0.16, 0.35)
GAM-S (smoothing splines)	0.27 (0.18, 0.37)	0.35 (0.24, 0.46)	0.26 (0.16, 0.37)

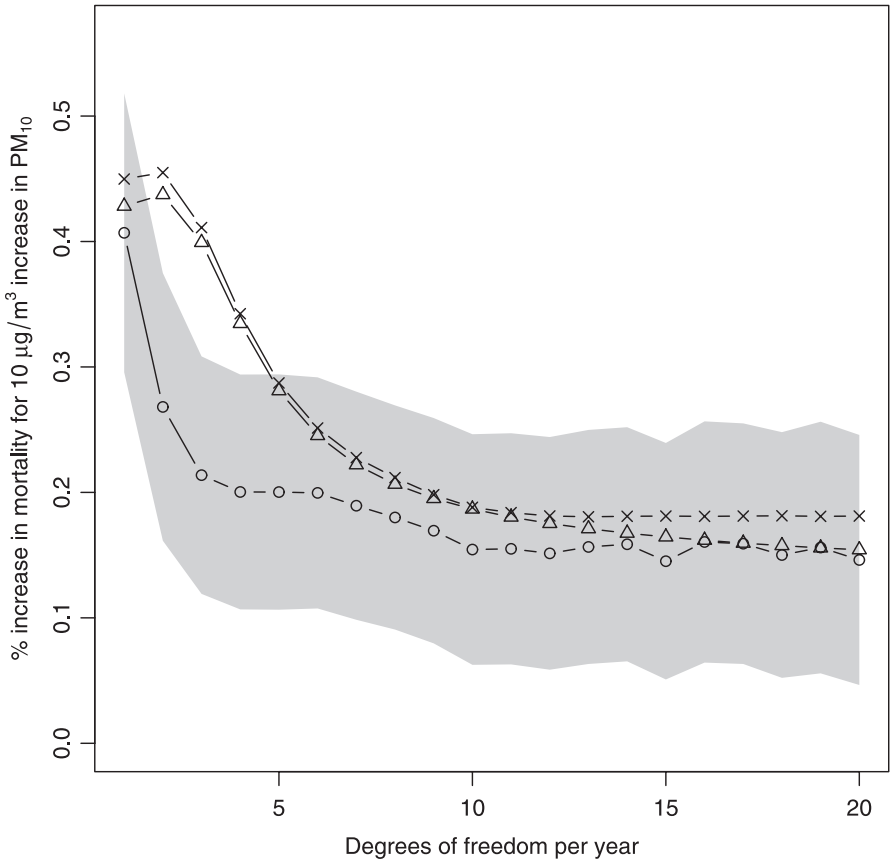

Fig. 3. Sensitivity analysis of the national average estimate of the percentage increase in mortality for an increase in PM_{10} of 10 $\mu g\,m^{-3}$ at lag 1: city-specific estimates were obtained from 100 US cities using data for the years 1987–2000 and the estimates were combined by using a hierarchical normal model (○, GLM-NS; △, GAM-R; ×, GAM-S; ■, 95% posterior intervals for the estimates obtained by using GLM-NS)

fixed number of degrees of freeedom per year for all the cities. Fig. 3 shows the change in the national average estimate as df is varied. When using natural splines, the estimates appear to stabilize after df = 9 per year at around a 0.15% increase in mortality with an increase of 10 $\mu g\,m^{-3}$ in PM_{10} at lag 1. The estimates that were obtained by using smoothing splines also appear to stabilize, but at a higher value. The estimates that were obtained by using penalized splines are very close to the smoothing spline estimates up to approximately df = 12 per year, after which the penalized spline estimates decrease slightly.

5. Discussion

We have developed a framework for quantifying and characterizing model uncertainty in multicity time series studies of air pollution and mortality. The complexity of the time series data requires the application of sophisticated statistical models that are capable of estimating relatively small effects. Furthermore, these effects have important policy implications, making a critical evaluation of the diverse modelling approaches that have been proposed in the literature an important task.

We have conducted a simulation study to compare commonly used approaches to adjusting for seasonal and long-term trends in air pollution epidemiology under a variety of realistic scenarios of confounding. The simulations quantify the average bias and standard error that are associated with each of the different modelling approaches. In addition to the simulation study we have applied all the methods to the NMMAPS database, the largest publicly available database containing time series data of air pollution and mortality. Our analysis of the NMMAPS data is important because it demonstrates that the national average estimates of the effect of PM_{10} at lag 1 are robust to different model selection criteria and smoothing methods. The results that are presented here strengthen recent findings from multicity time series studies regarding the effects of short-term increases in air pollution on daily mortality.

We have focused on the smooth function of time that is used to control for seasonal and long-term trends in mortality. The different approaches to representing the smooth function and to specifying its smoothness have varying effects on the bias and variance of the estimates depending on how the methods are combined and on the concurvity in the data. When using data-driven methods to specify the smoothness, higher concurvity leads to more biased estimates as does using penalized splines over natural splines, although the effect of concurvity is far greater.

Our results show that both fully parametric and nonparametric methods perform well, with neither preferred. A sensitivity analysis from the simulation study indicates that neither the natural spline nor the penalized spline approach produces any systematic bias in the estimates of the log-relative-rate β. However, that is not to say that the two approaches are equivalent; the data analysis must be tuned to the specific approach. The results of Rice (1986) and Speckman (1988) suggest that, with a nonparametric approach (such as penalized splines), we must use a df that is *not* optimal for predicting mortality to obtain an estimate of β with an acceptable rate of convergence for the bias. The simulation study in Section 3 confirms this notion in that we need to use a larger df to achieve the same average bias as the corresponding estimate obtained via natural splines (see for example Fig. 2). Therefore, the automatic use of criteria such as generalized cross-validation or AIC for selecting df could be potentially misleading (particularly with high concurvity) since they are designed to choose the df that will lead to optimal prediction of the mortality series, not necessarily to accurate estimation of β.

For parametric models (with natural splines), Dominici *et al.* (2004) showed that we must use a df that is at least as large as that needed to predict the *pollution series* best. They suggested using a procedure such as generalized cross-validation to estimate this df and then to use the bootstrap to minimize an estimate of the mean-squared error for $\hat{\beta}$. Our simplified version (GCV-PM_{10}) of their approach performs very well in the simulations and produces estimates of β that are nearly unbiased under all the scenarios, even under high concurvity.

The failure of BIC to produce competitive estimates of β, although dramatic, is not of concern in assessing the relationship between air pollution and health because it has generally not been applied. Although it is sometimes used to provide an approximate Bayes posterior (relative) probability for each df, our modelling set-up is far from that considered by Schwarz (1978). That is, as $n \to \infty$, we also have that the dimension of the model tends to ∞, which can lead BIC to choose the wrong model (Stone, 1979; Berger *et al.*, 2003; Hansen and Yu, 2003). The use of BIC in this setting, for example, in conjunction with Bayesian model averaging, requires further exploration.

Under moderate concurvity, AIC produces estimates of β with relatively small bias. Shibata (1976) demonstrated for autoregressive time series models that AIC has the potential to select increasingly larger models as the sample size increases (see also Ripley (1996)), a feature that is perhaps desirable here. Stone (1979) also showed that, in certain situations where the dimension

of the model tends to ∞, AIC can choose the right model as $n \to \infty$. However, it is important to note that using AIC to select the df that best predicts mortality still may not lead to the best estimate of β in this setting. For example, in Table 2, we see that, when g is rougher than f, the estimates that are selected by AIC are much more biased than when g is smoother than f.

Selecting the degrees of freedom for the smooth function of time by minimizing autocorrelation in the residuals is a heuristic approach that is widely used in the air pollution and health literature. Schwartz (1994b) suggested that the presence of residual autocorrelation may lead to underestimation of standard errors and, as a result, biased hypothesis tests of the pollutant variable coefficient; minimizing such autocorrelation would seem a natural goal. Although underestimation of standard errors can lead to possibly incorrect inferences about the city-specific coefficients, Daniels *et al.* (2004) showed that in a multicity context the underestimation of the city-specific standard errors would have to be severe (or the number of cities very small) to result in a substantial change in the national average (pooled) estimate.

Our simulation study indicates that inducing some residual (negative) autocorrelation may be necessary to reduce the bias in estimates of the pollution coefficient $\hat{\beta}$. Fig. 2 shows that increasing df tends to decrease the bias in the pollution coefficient estimates while slightly increasing the variability of these estimates. Table 2 indicates that, with penalized splines, using the true df may not be sufficient to reduce the bias in $\hat{\beta}$. Generally, undersmoothing the data (i.e. increasing df for the smooth function of time) induces residual autocorrelation at a number of lags.

The conclusion that residual autocorrelation may be necessary to control for confounding bias emphasizes the importance of distinguishing between *model uncertainty* and *adjustment uncertainty*. When addressing model uncertainty, we select the covariates that best explain the variability in the response which, in our setting, would require selecting df to obtain white noise in the residuals. With adjustment uncertainty, we select the covariates that minimize confounding bias in the exposure effect estimate. Previous contributions in semiparametric regression (Speckman, 1988; Dominici *et al.*, 2004) have shown that, if the goal of inference is confounding adjustment, the model should include all the covariates that are needed to explain variation in the exposure of interest, not the outcome. Therefore, in our setting, we need to select enough degrees of freedom for the smooth function of time to explain the variation in air pollution. This selected df might be smaller or larger than the optimal one that is needed to explain variation in the response, thus leaving autocorrelation in the residuals.

Also of concern is the application of the minimum PACF procedure to data sets with regular patterns of missing data. Although the NMMAPS analysis in Section 4 indicates that the effects of the missing data are not profound, it nevertheless seems inappropriate to apply this procedure for those data.

All our conclusions from the simulation study are based on assuming a true $\beta = 0$. Although our results would generalize in a standard linear regression framework to situations where $\beta \neq 0$, the use of a non-identity link function here precludes such generalization. The performance of all the estimation methods for $\beta \neq 0$ merits exploration. However, with time series models for air pollution and mortality an important concern is distinguishing correctly between a very small, but non-zero, effect and a true zero effect. Hence, in this paper we have concentrated on the scenario where the true β is 0.

Although incorporating a smooth function of time is a widely used method to control for seasonal patterns, it is by no means the only option. Case–crossover analyses (Navidi, 1998) have also been applied to the US data and represent an entirely different approach to controlling for confounding by season (Schwartz *et al.*, 2003; Schwartz, 2004). The results in those studies were qualitatively similar to those which were obtained here for the effect of PM_{10} at lag 1, although the estimates that were obtained in Schwartz *et al.* (2003) were slightly higher.

Of course, the case–crossover analyses also face challenging model choice questions such as choosing the 'window' for selecting referent cases or controls. Nevertheless, the analyses are relevant because they further reinforce the notion that results from multicity time series studies are robust to alternative methodologies and data analytic approaches.

Acknowledgements

This research was supported in part by National Institutes of Health–National Heart, Lung, and Blood Institute grant T32HL07024, National Institute of Environmental Health Sciences grant R01ES012054, the National Institute of Environmental Health Sciences Center for Urban Environmental Health (P30ES03819). Research that is described in this paper was partially supported by a contract and grant from the Health Effects Institute, an organization that is jointly funded by the US Environmental Protection Agency and automotive manufacturers. The contents of this paper do not necessarily reflect the views and policies of the Health Effects Institute; nor do they necessarily reflect the views and policies of the Environmental Protection Agency, nor motor vehicle or engine manufacturers.

Appendix A: Details of representations for *f*

Natural splines are piecewise cubic polynomials that are defined on a grid of knot locations spanning the range of the data. The function itself, as well as its second derivative, is continuous on the entire range of the data and the function is restricted to be linear beyond the end points. The smoothness of a natural spline fit is controlled by the number of knots that are used. Fewer knots represent smoother fits whereas n knots (where n is the sample size) will lead to interpolation of the data. The knot locations are often chosen to be at regressor values associated with equally spaced quantiles but could, in principle, be anywhere.

Penalized splines can provide a more flexible way to model non-linear relationships. They have been presented in the literature in various ways and we use the general definition, $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{B}(x)$, where

$$\hat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} \left[\sum_{i=1}^{n} \{y_i - \boldsymbol{\eta}^{\mathrm{T}}\mathbf{B}(x_i)\}^2 + \alpha \boldsymbol{\eta}^{\mathrm{T}}\mathbf{H}\boldsymbol{\eta} \right].$$

$\mathbf{B}(x)$ is a spline basis matrix (evaluated at the point x), α is a penalty (smoothing) parameter and $\mathbf{H}$ is a penalty matrix.

Versions of penalized splines essentially boil down to different specifications of the spline basis matrix $\mathbf{B}$ and the form of the penalty $\mathbf{H}$. A common approach constructs a natural spline or B-spline basis using a large number of knots (far more than are generally considered necessary) and then shrinks the coefficients to reduce the effective degrees of freedom and increases smoothness in the overall function estimate (Marx and Eilers, 1998; Wood, 2000). The amount of smoothness in the estimated curve (i.e. shrinking of the coefficients) is controlled by α. As $\alpha \downarrow 0$, the amount of smoothing decreases and the estimated curve approaches the full parametric fit. As $\alpha \uparrow \infty$, the amount of smoothing increases and the estimated curve approaches a polynomial function.

The most extreme approach to knot placement in the penalized spline framework is to place the maximum number of knots possible, i.e. one knot at every data point. The resulting fit is then called a smoothing spline. Time series data are typically regularly spaced and the smoothing spline scheme leads to n equally spaced knots along the time period of the data set. Since smoothing splines can be considered a special case of penalized splines (Ruppert *et al*., 2003), we expect that results which are obtained by using smoothing splines and penalized splines would be very similar, except perhaps in the case of penalized splines where too few knots are used (see for example the discussion in Eilers and Marx (1996)).

The complexity of a spline basis representation can be measured by its degrees of freedom. Since the previously mentioned approaches are linear, they can be represented by the $n \times n$ smoother matrix which maps the observed data to the smooth predicted values. The *effective degrees of freedom* are computed by the trace of the *smoother matrix* (Buja *et al*., 1989; Hastie and Tibshirani, 1990). For fully parametric fits such as those using natural splines, this trace equals the number of estimated parameters in the model.

A.1. Estimation of β

For the purposes of this section, we shall take the more simplified version of model (1), focusing on the estimation of β and the smooth function of time f. Using matrix notation, we can rewrite model (1) as

$$\begin{aligned} \mathbf{Y} &\sim \text{Poisson}(\boldsymbol{\mu}), \\ \log(\boldsymbol{\mu}) &= X\boldsymbol{\beta} + \mathbf{f} \end{aligned} \tag{5}$$

where $\mathbf{Y} = y_1, \ldots, y_n$, $\mathbf{f}$ is the function f evaluated at $t = 1, \ldots, n$ and X is the $n \times 2$ design matrix containing a column of 1s and the pollution time series $x_1, \ldots, x_n$.

Given one of the spline bases that are described in Section 2.1, we can rewrite model (5) as

$$\log(\boldsymbol{\mu}) = X\boldsymbol{\beta} + B\boldsymbol{\gamma}$$

where B is the $n \times d$ matrix of d basis functions and $\boldsymbol{\gamma}$ is a d-vector of coefficients. The number of columns of the basis matrix B will be different depending on whether natural splines, penalized splines or smoothing splines are used.

We use iteratively reweighted least squares to fit model (5) by using natural splines. Let W be the $n \times n$ (diagonal) weight matrix and $\mathbf{z}$ the working response from the last iteration of the iteratively reweighted least squares algorithm. Let X^* be the complete design matrix, i.e. $X^* = (X|B)$. Using a generalized linear model procedure with natural cubic splines, we can estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ simultaneously as

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_{\text{ns}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = (X^{*\text{T}} W X^*)^{-1} X^{*\text{T}} W \mathbf{z}.$$

For penalized splines, we first need to construct the smoother matrix for the nonparametric part of the model. Given a value for the smoothing parameter α and a fixed (symmetric) penalty matrix H, the smoother matrix for $\mathbf{f}$ is

$$S = B(B^{\text{T}}B + \alpha H)^{-1} B^{\text{T}}$$

and the estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{\text{ps}} = (X^{\text{T}} W (I - S) X)^{-1} X^{\text{T}} W (I - S) \mathbf{z}.$$

Rice (1986) and Speckman (1988) both showed that, whereas the variance of $\hat{\boldsymbol{\beta}}_{\text{ps}}$ converges at the standard parametric rate for $n \to \infty$, the bias converges to 0 at the much slower nonparametric rate. The slow convergence of the bias comes from the fact that the smoother matrix S is not a true projection, unlike the hat matrix in parametric regression (Speckman, 1988). The performance of both $\hat{\boldsymbol{\beta}}_{\text{ns}}$ and $\hat{\boldsymbol{\beta}}_{\text{ps}}$ by using various model selection criteria was illustrated in Sections 3 and 4.

Speckman (1988) described an alternative estimator for $\boldsymbol{\beta}$ for which the bias and variance both converge at the usual parametric rate. For S symmetric, the modified estimator is

$$\hat{\boldsymbol{\beta}}^*_{\text{ps}} = (X^{\text{T}} W (I - S)^2 X)^{-1} X^{\text{T}} W (I - S)^2 \mathbf{z}.$$

If we let $\tilde{X} = (I - S)X$ and $\tilde{\mathbf{z}} = (I - S)\mathbf{z}$, then the modified estimator can be written as

$$\hat{\boldsymbol{\beta}}^*_{\text{ps}} = (\tilde{X}^{\text{T}} W \tilde{X})^{-1} \tilde{X}^{\text{T}} W \tilde{\mathbf{z}},$$

which is the estimate that we might obtain from a regression of $\tilde{\mathbf{z}}$ on $\tilde{X}$. Hence, this modified estimator has the form of a regression of partial residuals.

A simple calculation shows that estimating $\boldsymbol{\beta}$ by using $\hat{\boldsymbol{\beta}}^*_{\text{ps}}$ is equivalent to estimating $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}_{\text{ps}}$ but with the modified smoother matrix

$$\begin{aligned} \tilde{S} &= I - (I - S)^2 \\ &= S(2I - S). \end{aligned}$$

Buja *et al.* (1989) showed that $\text{tr}(\tilde{S}) \geqslant \text{tr}(S)$ and hence, to obtain an estimate of $\boldsymbol{\beta}$ for which the bias and variance converge at the parametric rate, we must implicitly use an undersmoothed estimate of $\mathbf{f}$. If Λ

is the diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_n$ for the smoother matrix S, a simple calculation reveals that

$$\begin{aligned}\mathrm{tr}(\tilde{S}) &= \mathrm{tr}\{S(2I - S)\} \\ &= \mathrm{tr}\{\Lambda(2I - \Lambda)\} \\ &= \sum_{i=1}^{n} 2\lambda_i - \lambda_i^2 \\ &= \mathrm{tr}(S) + \sum_{i=1}^{n} \lambda_i(1 - \lambda_i). \end{aligned} \tag{6}$$

The quantity in the summation can be interpreted as the extra degrees of freedom that are required for the modified estimate $\hat{\beta}^*_{\mathrm{ps}}$, i.e. the amount of undersmoothing that is required. It is important to note that the extra degrees of freedom in equation (6) may be small and, furthermore, using $\hat{\beta}^*_{\mathrm{ps}}$ only provides the same rate of convergence for the bias as using $\hat{\beta}_{\mathrm{ns}}$. For a fixed n the two estimates may be quite different.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csáki), pp. 267–281. Budapest: Akadémiai Kiadó.

Bell, M. L., Samet, J. M. and Dominici, F. (2004) Time-series studies of particulate matter. *A. Rev. Publ. Hlth*, **25**, 247–280.

Berger, J. O., Ghosh, J. K. and Mukhopadhyay, N. (2003) Approximations and consistency of Bayes factors as model dimension grows. *J. Statist. Planng Inf.*, **112**, 241–258.

Brockwell, P. J. and Davis, R. A. (2002) *Introduction to Time Series and Forecasting*, 2nd edn. New York: Springer.

Buja, A., Hastie, T. and Tibshirani, R. (1989) Linear smoothers and additive models. *Ann. Statist.*, **17**, 453–555.

Burnett, R. T., Cakmak, S. and Brook, J. R. (1998) The effect of the urban ambient air pollution mix on daily mortality rates in 11 Canadian cities. *Can. J. Publ. Hlth*, **89**, 152–156.

Burnett, R. T. and Goldberg, M. S. (2003) Size-fractionated particulate mass and daily mortality in eight Canadian cities. In *Revised Analyses of Time-series Studies of Air Pollution and Health*, pp. 85–89. Cambridge: Health Effects Institute.

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

Curriero, F. C., Heiner, K. S., Samet, J. M., Zeger, S. L., Strug, L. and Patz, J. A. (2002) Temperature and mortality in 11 cities of the Eastern United States. *Am. J. Epidem.*, **155**, 80–87.

Daniels, M. J., Dominici, F., Samet, J. M. and Zeger, S. L. (2000) Estimating PM_{10}-mortality dose-response curves and threshold levels: an analysis of daily time-series for the 20 largest US cities. *Am. J. Epidem.*, **152**, 397–412.

Daniels, M. J., Dominici, F. and Zeger, S. L. (2004) Underestimation of standard errors in multi-site time series studies. *Epidemiology*, **15**, 57–62.

Dockery, D. W. and Pope, C. A. (1996) Epidemiology of acute health effects: summary of time-series studies. In *Particles in Our Air* (eds R. Wilson and J. Spengler), pp. 123–147. Boston: Harvard University Press.

Dominici, F., McDermott, A., Daniels, M., Zeger, S. L. and Samet, J. M. (2003) Mortality among residents of 90 cities. In *Revised Analyses of Time-series Studies of Air Pollution and Health*, pp. 9–24. Cambridge: Health Effects Institute.

Dominici, F., McDermott, A. and Hastie, T. (2004) Improved semiparametric time series models of air pollution and mortality. *J. Am. Statist. Ass.*, **99**, 938–948.

Dominici, F., McDermott, A., Zeger, S. L. and Samet, J. M. (2002) On the use of generalized additive models in time-series studies of air pollution and health. *Am. J. Epidem.*, **156**, 193–203.

Dominici, F., Samet, J. M. and Zeger, S. L. (2000) Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy (with discussion). *J. R. Statist. Soc.* A, **163**, 263–302.

Donnell, D. J., Buja, A. and Stuetzle, W. (1994) Analysis of additive dependencies and concurvities using smallest additive principal components. *Ann. Statist.*, **22**, 1635–1668.

Eilers, P. H. C. and Marx, B. D. (1996) Flexible smoothing using *B*-splines and penalized likelihood (with discussion). *Statist. Sci.*, **11**, 89–121.

Environmental Protection Agency (1996) Air quality criteria for particulate matter. *Report EPA/600/P-95/001aF*. Office of Research and Development, Environmental Protection Agency, Washington DC.

Environmental Protection Agency (2003) Air quality criteria for particulate matter (fourth external review draft). *Reports EPA/600/P-99/002aD and EPA/600/P-99/002bD*. Office of Research and Development, National Center for Environmental Assessment, Research Triangle Park.

Everson, P. J. and Morris, C. N. (2000a) Inference for multivariate normal hierarchical models. *J. R. Statist. Soc.* B, **62**, 399–412.

Everson, P. J. and Morris, C. N. (2000b) Simulation from Wishart distributions with eigenvalue constraints. *J. Computnl Graph. Statist.*, **9**, 380–389.
Goldberg, M. S., Burnett, R. T., Brook, J., Bailar, J. C. I., Valois, M.-F. and Vincent, R. (2001) Associations between daily cause-specific mortality and concentrations of ground-level ozone in Montreal, Quebec. *Am. J. Epidem.*, **154**, 817–826.
Goldberg, M. S., Burnett, R. T. and Stieb, D. (2003) A review of time-series studies used to evaluate the short-term effects of air pollution on human health. *Rev. Environ. Hlth*, **18**, 269–303.
Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. London: Chapman and Hall.
Gu, C. (2002) *Smoothing Spline ANOVA Models*. New York: Springer.
Hansen, M. and Yu, B. (2003) Minimum description length model selection criteria for generalized linear models. *IMS Lect. Notes*, **40**, 145–164.
Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. New York: Chapman and Hall.
Health Effects Institute (2003) *Revised Analyses of Time-series Studies of Air Pollution and Health: Special Report*. Boston: Health Effects Institute.
Katsouyanni, K., Toulomi, 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*, **12**, 521–531.
Kelsall, J. E., Samet, J. M., Zeger, S. L. and Xu, J. (1997) Air pollution and mortality in Philadelphia, 1974–1988. *Am. J. Epidem.*, **146**, 750–762.
Marx, B. D. and Eilers, P. H. C. (1998) Direct generalized additive modeling with penalized likelihood. *Computnl Statist. Data Anal.*, **28**, 193–209.
Navidi, W. (1998) Bidirectional case–crossover designs for exposures with time trends. *Biometrics*, **54**, 596–605.
Peng, R. D., Dominici, F., Pastor-Barriuso, R., Zeger, S. L. and Samet, J. M. (2005) Seasonal analyses of air pollution and mortality in 100 US cities. *Am. J. Epidem.*, **161**, 585–594.
Peng, R. D. and Welty, L. J. (2004) The NMMAPSdata package. *R News*, **4**, 10–14.
Pope, C. A., Dockery, D. W. and Schwartz, J. (1995) Review of epidemiological evidence of health effects of particulate air pollution. *Inhaln Toxicol.*, **47**, 1–18.
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.
R Development Core Team (2003) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rice, J. (1986) Convergence rates for partially splined models. *Statist. Probab. Lett.*, **4**, 203–208.
Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.
Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. Cambridge: Cambridge University Press.
Samet, J. M., Zeger, S. L., Dominici, F., Curriero, F., Coursac, I., Dockery, D., Schwartz, J. and Zanobetti, A. (2000a) *The National Morbidity, Mortality, and Air Pollution Study*, part II, *Morbidity and Mortality from Air Pollution in the United States*. Cambridge: Health Effects Institute.
Samet, J. M., Zeger, S. L., Dominici, F., Dockery, D. and Schwartz, J. (2000b) *The National Morbidity, Mortality, and Air Pollution Study*, part I, *Methods and Methodological Issues*. Cambridge: Health Effects Institute.
Samet, J. M., Zeger, S. L., Kelsall, J., Xu, J. and Kalkstein, L. (1998) Does weather confound or modify the association of particulate air pollution with mortality? *Environ. Res.* A, **77**, 9–19.
Samoli, E., Schwartz, J., Wojtyniak, B., Touloumi, G., Spix, C., Balducci, F. and Medina, S. (2002) Investigating regional differences in short-term effects of air pollution on daily mortality in the APHEA project: a sensitivity analysis for controlling long-term trends and seasonality. *Environ. Hlth Perspect.*, **109**, 349–353.
Samoli, E., Touloumi, G., Zanobetti, A., Le Tertre, A., Schindler, C., Atkinson, R., Vonk, J., Rossi, G., Saez, M., Rabczenko, D., Schwartz, J. and Katsouyanni, K. (2003) Investigating the dose-response relation between air pollution and total mortality in the APHEA-2 multicity project. *Occupnl Environ. Med.*, **60**, 977–982.
Schwartz, J. (1994a) Nonparametric smoothing in the analysis of air pollution and respiratory illness. *Can. J. Statist.*, **22**, 471–488.
Schwartz, J. (1994b) Total suspended particulate matter and daily mortality in Cincinnati, Ohio. *Environ. Hlth Perspect.*, **102**, 186–189.
Schwartz, J. (2000) The distributed lag between air pollution and daily deaths. *Epidemiology*, **11**, 320–326.
Schwartz, J. (2004) Is the association of airborne particles with daily deaths confounded by gaseous air pollutants?: an approach to control by matching. *Environ. Hlth Perspect.*, **112**, 557–561.
Schwartz, J., Zanobetti, A. and Bateson, T. (2003) Morbidity and mortality among elderly residents of cities with daily PM measurements. In *Revised Analyses of Time-series Studies of Air Pollution and Health*, pp. 25–58. Cambridge: Health Effects Institute.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **5**, 461–464.
Shibata, R. (1976) Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, **63**, 117–126.
Speckman, P. (1988) Kernel smoothing in partial linear models. *J. R. Statist. Soc.* B, **50**, 413–436.

Stone, M. (1979) Comments on model selection criteria of Akaike and Schwarz. *J. R. Statist. Soc.* B, **41**, 276–278.
Touloumi, G., Atkinson, R., Le Tertre, A., Samoli, E., Schwartz, J., Schindler, C., Vonk, J., Rossi, G., Saez, M., Rabszenko, D. and Katsouyanni, K. (2004) Analysis of health outcome time series data in epidemiological studies. *Environmetrics*, **15**, 101–117.
Welty, L. J. and Zeger, S. L. (2005) Are the acute effects of PM_{10} on mortality in NMMAPS the result of inadequate control for weather and season?: a sensitivity analysis using flexible distributed lag models. *Am. J. Epidem.*, **162**, 80–88.
Wood, S. N. (2000) Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Statist. Soc.* B, **62**, 413–428.
Wood, S. N. (2001) mgcv: GAMs and generalized ridge regression for R. *R News*, **1**, 20–25.

Comments on the paper by Peng, Dominici and Louis

Joel Schwartz (*Harvard School of Public Health, Boston*)

Time series analysis was developed for Gaussian data, and two main threads exist to deal with potential confounding by unmeasured covariates that vary over time. In the approach of Box and Jenkins (1976), the dependent variable is filtered, usually to white noise. The independent variables are then filtered using the same autoregressive integrated moving average filter as was found adequate to 'prewhiten' the dependent variable, and then the two filtered series are regressed. In the approach of Haugh and Box (1977), each series is separately prefiltered using potentially different autoregressive integrated moving average filters, which are the ones that best whiten each series. Again, the filtered series are regressed against each other. The former approach has more of the feeling of a multiple-regression analysis of y against x_1 and x_2, since the same x_2, in this case a smooth function of time, is removed from x_1 and y. The latter approach was developed to allow better identification of the specific lag between exposure and response, when the exposure variable exhibits serial correlation.

Mortality count data are Poisson distributed, and these approaches are not available. The usual approach has been to use a Poisson regression, and to put the time filter in the model as $f(t)$, as noted by the authors. The question of how many degrees of freedom to use for time is hence analogous to the question of how much prefiltering to use. Peng and colleagues have explored this question, and the degree of bias that we might expect, using simulation analyses, and then applied these lessons to a large multicity database of PM_{10} and mortality. They are to be commended for this work. Several features of it deserve further discussion.

First, the finding in the simulations that more degrees of freedom per year are needed for penalized splines than for natural splines fits in well with our understanding of the need for undersmoothing. Penalized splines are more flexible than natural splines, and the likelihood of bias without undersmoothing is therefore stronger.

More important, however, is the difference between the figures showing the bias in the simulation studies as a function of degrees of freedom, and the comparable figures for the National Morbidity, Mortality and Air Pollution Study (NMMAPS) reanalysis. For natural splines, the bias seems to drop to 0 by 4 degrees of freedom per year with high or low concurvity, and smoother or rougher $g(t)$ relative to $f(t)$. In contrast, the natural spline model for the NMMAPS does not asymptote out until 10 degrees of freedom per year. Why is this?

I suggest that it is because the effects of particulate air pollution are distributed over multiple days, and air pollution is serially correlated. If pollution at lags 0, 1 and 2 is associated with the risk of mortality, and the model contains only exposure at lag 1, that variable will capture some of the effects of the collinear exposures at lag 0 and lag 1. This is why Haugh recommended prefiltering the exposure series to white noise if one wished to identify the lag relationship with outcome. I believe that this is also why the automatic span selection rule of the authors' GCV-PM_{10} strategy works better than the rest: it effectively chooses the degrees of freedom that are necessary to whiten the exposure series.

The idea that the effect of exposure to air pollution today is felt not merely tomorrow, but on multiple subsequent days, has support both from prior work on time series and from more focused health studies. Time domain (Schwartz, 2000) and frequency domain (Zeger *et al.*, 1999) regressions suggested that the effect size estimates for particles doubles as we move to longer timescales. More recently, Zanobetti *et al.* (2000, 2002) examined this question by respectively using smoothed distributed lag models and unconstrained distributed lag models containing multiple lags of the exposure simultaneously. Those give inefficient estimates of effects at individual lags, but unbiased estimates of the sum of coefficients across lags, which index the overall effect. A theoretical construct for these results has been proposed (Schwartz, 2001), whereby exposure induces the transition of a subject from a low risk to a frail pool. Until they

recover, and leave that pool, they are at increased risk of dying from a variety of stimuli. For example, exposure to particles has been shown to double the volume of the lung with pneumonia in an animal model. Once a more severe pneumonia has been induced, the patient would be at increased risk of dying for the next several weeks, until the natural recovery processes remove the surviving patients from the high risk pool.

If the additional degrees of freedom that are required to stabilize the estimated effect at lag 1 in the NMMAPS analysis are the result of confounding bias by other lags of exposure, this has implications for other studies where daily exposure data (instead of sixth-day exposure data in the NMMAPS) are available. Those studies can examine the distributed lag effects of particles. In an unconstrained distributed lag model, each other lag is simultaneously in the regression, suggesting that additional filtering should not be necessary to remove bias. Hence the 4 degrees of freedom per year results from the simulation studies may be closer to the mark for what is required in such analyses.

To illustrate this, I have done a simulation. I used the PM_{10}-data from Chicago from 1988 to 1993, to reflect the true serial correlation that is found in such data. I then assumed a true distributed lag between exposure and the log-relative-risk of death that was highest in the first few days, but stretched out 45 days, as suggested by the work of Zanobetti and co-workers (Schwartz, 2001). The summed coefficients in this model were 0.00130. From this I simulated a Poisson count for the 6 years of data with a sinusoidal seasonal pattern. I then fitted an unconstrained distributed lag model and summed the coefficients to obtain the estimated overall effect. I repeated my simulation 300 times, and the results are shown in Fig. 4. A value of 5–7 degrees of freedom per year seems to give unbiased results, with larger degrees of freedom per year introducing a negative bias, perhaps because they start to remove fluctuations of the order of the distributed lag. This makes the choice of degrees of freedom more challenging than in the case of estimating a single lag. We can perform the sensitivity analyses that were suggested by Peng and co-workers, but note the problem, or we can rely on what we know of the sources of variation of fluctuations in mortality and air pollution on different timescales to provide some guidance.

It is worth noting that the question of how much filtering to do is not entirely an analytic one. We remove fluctuations of longer timescales and keep short-term fluctuations because we believe that there are omitted confounders that are related to both outcome and exposure at longer wavelengths, but not at very short wavelengths. We have some information about the boundary, because we know quite considerably about the sources of fluctuations in air pollution, and a fair amount about fluctuations in mortality. An omitted

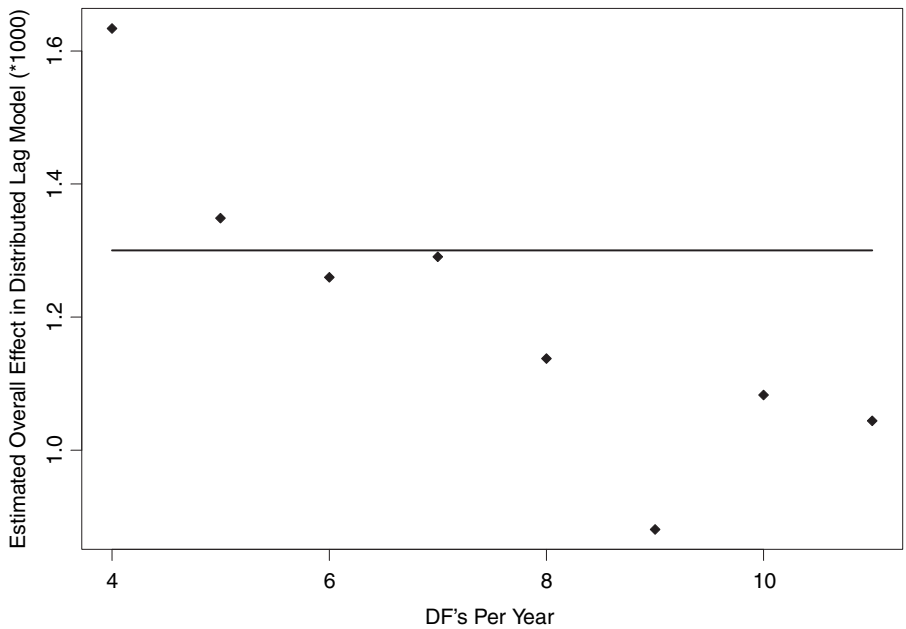

Fig. 4. Estimated overall distributed lag effects

confounder is causally related to air pollution and death. Is there a common aetiology for variations in particles and death on subseasonal timescales? Temperature and other pollutants of course do covary with particles, but these are dealt with in the NMMAPS analyses. What else could confound? Here we can draw on our considerable knowledge of what causes short-term temporal variations in air quality.

The principal determinants of variation in air pollution concentrations are day of the week changes in emissions (but the day of the week is controlled in these studies), day-to-day changes in the height of the mixing layer and changes in the trajectory describing where the air currently over the city came from. The mixing height determines the location of the layer of inversion that traps pollution from rising further. This can vary widely, resulting in substantial changes in concentrations. The back-trajectory of the air mass determines whether the air came from a more or less heavily emitting region before arriving at the city under study. Again, this can cause substantial variations in concentration.

Given this knowledge, we can ask how does the height of the mixing layer or the back-trajectory of the air mass cause variations in other predictors of the risk of mortality that are not on the causal pathway for particulate air pollution or already controlled? It is difficult to believe that diet changes with back-trajectory for example. How do people in Philadelphia know whether their air was in Ohio or in Ontario 2 days ago? Hence confounding by very short-term phenomena seems implausible because the known predictors of short-term variations in air pollution concentrations are either weather terms that are already controlled in the model or processes that are unlikely to cause variations in other health predictors. As we move past a few days, the issue becomes less clear, but the sources of air pollution variation are predominantly the same.

Respiratory epidemics generally last for a few weeks and are often cited as a potential confounder for timescales of that wavelength. Indeed Peng and coworkers argue that many degrees of freedom are required for $f(t)$ to pick up those epidemics. However, although epidemics peak in the winter, within that season, their occurrence does not covary with PM_{10}-levels. Braga *et al.* (2000) used hospital admissions for pneumonia to identify respiratory epidemic periods in each year in multiple cities. They then examined the effect of PM_{10} on daily deaths in models with 4 degrees of freedom per year seasonal control, and no control for epidemics, compared with models that fit a separate 6 degrees of freedom curve for each epidemic period, allowing the length, height and shape of the rise and fall of mortality with the epidemic to vary. Multiple epidemics per year were often found. The PM_{10}-coefficient was unchanged after this control. Similarly, I previously repeated multiple analyses excluding all days above the 97th-percentile of temperature, and I found no change in PM_{10}-coefficients. So heat waves, which may not be captured by the weather parameters, also seem unlikely as confounders. These arguments, as well as the simulation results, suggest that, when multiple pollutant lags are included in a model simultaneously, temporal control of the order of 4–6 degrees of freedom per year appears adequate.

Finally, the existence of daily pollution data also raises the need to control for serial correlation in the data (which is generally absent with only sixth-day measurements). If daily data are used, even when the degrees of freedom are chosen to minimize serial correlation, I have often found that some significant serial correlation remains. At high degrees of freedom per year, significant negative serial correlation is the rule. Although ignoring serial correlation will result in unbiased (but inefficient) estimates of the regression coefficients, it can result in biased estimates of the standard errors. Often, these are too small but, with negative serial correlation, it is possible that they are too large.

Peng and coworkers argue that, in large multicity studies, this has little effect on the estimate of overall effects. However, that was equally true for the misestimation of standard errors in the generalized additive model S-PLUS function: smaller within-city estimates lead to larger random effects in combining estimates across cities, and no noticeable difference in the overall estimate. Nevertheless, hundreds of studies were reanalysed as a result of the discovery of this problem, and every analyst will not necessarily have large numbers of locations to average over.

Several methods have been suggested to address this. The APHEA study has used autoregressive Poisson models as discussed by Brumback *et al.* (2000). Alternatively, Schwartz and Dockery (1992) suggested generalized estimating equations, treating each year as a replicate. Biggeri *et al.* (2005) have suggested season-specific models as reducing the complexity of the control for temporal confounding, with each year of each season again treated as replicates. This allows the replicates to be disjoint in time.

N. Reid (*University of Toronto*)

The authors are to be congratulated for pursuing the important question of model selection in the context of the time series models that are now widely used in the epidemiology literature. It is relatively easy to fit

very complex models, relatively quickly, but, as the 'generalized additive model crisis' showed, there can be unsuspected pitfalls by using available software.

The emphasis in the paper is on the interrelationship between smoothing and confounding, here attributed to 'concurvity'. Although the phrase '$f(\cdot)$ and $g(\cdot)$ are correlated at similar timescales' is an intuitive summary, on closer examination I found it difficult to understand what this meant. I think that it is more useful to concentrate on concurvity in the data, rather than a more theoretical version, and as long as the authors' 'GLM-NS' method is used this is simply multicollinearity, between the pollution measurements $x = (x_1, \ldots, x_n)$ and the linear representations that are used for the p confounders:

$$x \approx \sum_{j=1}^{p} a_j B_j \gamma_j,$$

using the notation of equation (5) in Appendix A, where each B_j is an $n \times d$ matrix of basis function evaluations for the jth confounder. Could linear regression diagnostics for collinearity be used in this case? Ramsay *et al.* (2003) suggested checking concurvity by calculating the correlation between the PM_{10}-measurements and the fitted values from a regression of PM_{10} on suspected confounders: do the authors have any experience with this in the context of their simulations?

The authors' definition is a little different, having $x \approx g(t) + \varepsilon$, which I think may be better for the more complex case of smoothing splines, where the representation of f_j is not explicit. In the simulations there is also a smooth function of temperature, $r(z_t)$, which may be of more importance in practice, and I wonder whether the authors have any results on the effect of varying the concurvity of r and q in a manner that is different from that of f and g, or is the problem 'symmetric' in this? It is reassuring that most of the model selection methods are fairly reliable, in the sense that is investigated here, and useful to know that the cross-validation criterion is essentially the most reliable; it has long been suspected that the widespread use of Akaike's information criterion for model selection, as opposed to prediction, was inappropriate. Are the results invariant to the true value of β, as intuition might suggest? In view of the combination of the city-specific estimates bias would seem to be much more important than variance, so there seems no reason to use smoothing splines nor the partial autocorrelation function.

A puzzle that is raised by Fig. 1, and I do not think is solved by the simulation study, is the empirical result that the pollution effect estimates are smaller when the confounding functions are estimated by natural cubic splines than by smoothing splines. I think that there is evidence in the literature that the estimated pollution effect is smaller still when estimated by a fixed and known parametric function, such as a fixed series of trigonometric terms. Can this simply be attributed to bias, or is something more subtle involved? We know that the maximum likelihood estimator of β may be inconsistent when the number of nuisance parameters increases with the sample size; it should in principle be possible to apply adjustments to the profile likelihood to lead to more accurate estimating equations but I do not know how complex this would be in practice.

There does seem to be growing evidence that there is a small but non-zero pollution effect on acute events, although the confounding with weather may not yet have been dealt with satisfactorily. Of possibly more public health interest are chronic effects, and research efforts on these will increasingly be important. The researchers of the Environmental Biostatistics and Epidemiology Group at Johns Hopkins University have played an important role in both science and public policy in moving this research agenda forward.

Authors' reply

We are grateful to Professor Reid and Professor Schwartz for their thoughtful discussion of our paper. They have raised some interesting points that require serious attention.

It is important to reiterate the principal conclusions of our work, all of which are highlighted by the discussants. First, our results show that, for single-site studies, what matters most is not how you smooth, but how much you smooth. Although estimates depend somewhat on the choice of basis (e.g. natural splines, penalized splines or smoothing splines) they primarily depend on the degrees of freedom that are used to control the amount of smoothing. Our simulations show that for the sorts of data we consider there is a range of degrees of freedom per year that is not sufficiently large to reduce bias. For natural splines this range is roughly 1–4 degrees of freedom per year whereas for penalized splines it is 1–9 degrees of freedom per year. The need for a larger number of degrees of freedom for penalized splines is explained by Rice (1986) and Speckman (1988), who documented the slow rate of decrease in the bias of parameter estimates in semiparametric models. Hence, when using nonparametric smoothers, to reduce bias in parameter estimates sufficiently, we must use more degrees of freedom per year than would be appropriate for a nonparametric function estimate (e.g. undersmooth). The need to undersmooth should not be taken

as a problem with nonparametric models, but as a fact of semiparametric life of which the practitioner should be keenly aware.

The decision about how much to smooth is by no means entirely analytic and we cannot rely solely on so-called automatic methods. Professor Reid raises the important point that the use of prediction-based criteria such as Akaike's information criterion is inappropriate for selecting an adjustment model. The simulation study supports this point, particularly in the scenarios where the degrees of freedom that are required to model the pollution series is larger than those required to model the mortality series (i.e. g rougher than f). In those scenarios, methods such as Akaike's information criterion and the partial autocorrelation function produce estimates of the pollution effect with generally larger bias relative to the GCV-PM_{10} method, which chooses the degrees of freedom on the basis of the pollution series rather than the mortality series. Concurvity is another factor that we must account for and Professor Reid makes the excellent suggestion to compute the empirical concurvity for each data set. Though we did not do so, the 'expected' degree of concurvity in our simulations is determined by the σ^2-parameter (our equation (3)) and it can be used to investigate concurvity–performance relations.

In the area of air pollution and health studies, sensitivity analysis is critical for uncovering and communicating the dependence of parameter estimates on adjustment procedures. In this regard Professor Schwartz has rightly noted that such a sensitivity analysis should not be conducted in a vacuum. We do understand to some degree what are the important time-varying confounders that lead to temporal variations in both air pollution levels and mortality counts.

Finally, it is clear that multisite studies provide more robust estimates of the short-term effect of air pollution than do single-site studies. Of course, the combined estimate is more stable than any of its component estimates but, as or more important, combining estimates from many sites has the potential to alleviate problems that are caused by misspecification of the site-specific models. In particular, if we underestimate the variance of an estimate at the site level (e.g. by failing to account sufficiently for serial correlation), the variance of the combined estimate based on a random-effects model can have relatively small bias. Daniels *et al*. (2004) illustrated that underestimation of the site-specific variances leads to larger estimates of the between-site variance (i.e. the heterogeneity parameter) in a random-effects model and vice versa. This trade-off between site-specific and between-site variance makes the combined estimate relatively insensitive to potential underestimation at the site level. However, empirical Bayes estimates of the site-specific log-relative-rates are sensitive to any underestimation of variances.

In this paper we did not address issues that are associated with accounting for potential confounding from temperature. All of the issues that we investigate with respect to modelling the smooth function of time should carry over to modelling temperature. However, Welty and Zeger (2005) showed using a rich family of models that sensitivity of the estimated PM_{10}-coefficient to the method that is used to control for temperature is of a smaller order than sensitivity to the amount of smoothing that is used to control for time. The national average estimates remain essentially unchanged when the temperature model is varied.

Professor Schwartz has raised the possibility that the effect of air pollution on mortality is spread over several days (a distributed lag) rather than induced by the exposure level on a single day, our single-lag model. He is almost certainly correct—it is very unlikely that the effect of an increase in air pollution would play out over only 1 day. However, our qualitative and quantitative comparisons and conclusions should apply equally well to distributed lag models. In practice, the use of single-lag models is largely a consequence of the '1 day in 6' sampling scheme for PM_{10} in most US cities. This sampling scheme prohibits the fitting of more complex distributed lag models unless complex multiple-imputation schemes are used. However, future availability of daily $PM_{2.5}$-data in several US cities and daily data on other pollutants will enable exploration of this issue on a national level.

References in the comments

Biggeri, A., Baccini, M., Bellini, P. and Terracini, B. (2005) Meta-analysis of the Italian studies of short-term effects of air pollution (MISA), 1990-99. *Int. J. Occup. Environ. Hlth*, **11**, 107–122.

Box, G. E. P. and Jenkins, G. M. (1976) *Time Series Analysis: Forecasting and Control*, revised edn. San Francisco: Holden–Day.

Braga, A. L., Zanobetti, A. and Schwartz, J. (2000) Do respiratory epidemics confound the association between air pollution and daily deaths? *Eur. Respir. J.*, **16**, 723–728.

Brumback, B. A., Ryan, L. M., Schwartz, J., Neas, L. M., Stark, P. C. and Burge, H. A. (2000) Transitional regression models with application to environmental time series. *J. Am. Statist. Ass.*, **95**, 16–28.

Daniels, M. J., Dominici, F. and Zeger, S. L. (2004) Understimation of standard errors in multi-site time series studies. *Epidemiology*, **15**, 57–62.

Haugh, L. D. and Box, G. E. P. (1977) Identification of dynamic regression (distributed lag) models connecting two time series. *J. Am. Statist. Ass.*, **72**, 121–130.
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.
Rice, J. (1986) Convergence rates for partially splined models. *Statist. Probab. Lett.*, **4**, 203–208.
Schwartz, J. (2000) Harvesting and long-term exposure effects in the relationship between air pollution and mortality. *Am. J. Epidem.*, **151**, 440–448.
Schwartz, J. (2001) Is there harvesting in the association of airborne particles with daily deaths and hospital admissions? *Epidemiology*, **12**, 55–61.
Schwartz, J. and Dockery, D. W. (1992) Particulate air pollution and daily mortality in Steubenville, Ohio. *Am. J. Epidem.*, **135**, 12–20.
Speckman, P. (1988) Kernel smoothing in partial linear models. *J. R. Statist. Soc.* B, **50**, 413–436.
Welty, L. J. and Zeger, S. L. (2005) Are the acute effects of PM_{10} on mortality in NMMAPS the result of inadequate control for weather and season?: a sensitivity analysis using flexible distributed lag models. *Am. J. Epidem.*, **162**, 80–88.
Zanobetti, A., Schwartz, J., Samoli, E., Gryparis, A., Touloumi, G., Atkinson, R., Le Tertre, A., Bobros, J., Celko, M., Goren, A., Forsberg, B., Michelozzi, P., Rabczenko, D., Ruiz, E. A. and Katsouyanni, K. (2002) The temporal pattern of mortality responses to air pollution. *Epidemiology*, **13**, 87–93.
Zanobetti, A., Wand, M. P., Schwartz, J. and Ryan, L. (2000) Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*, **1**, 279–292.
Zeger, S. L., Dominici, F. and Samet, J. (1999) Harvesting resistant estimates of air pollution effects on mortality. *Epidemiology*, **10**, 171–175.