6-1-2005

Model Choice in Time Series Studies of Air Pollution and Mortality

Roger D. Peng

Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, rpeng@jhsph.edu

Francesca Dominici

The Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, fdominic@jhsph.edu

Thomas A. Louis

Johns Hopkins University Bloomberg School of Public Health, Department of Biostatistics, tlouis@jhsph.edu

Suggested Citation

Peng, Roger D.; Dominici, Francesca; and Louis, Thomas A., "Model Choice in Time Series Studies of Air Pollution and Mortality" (June 2005). *Johns Hopkins University, Dept. of Biostatistics Working Papers*. Working Paper 55.
<http://biostats.bepress.com/jhubiostat/paper55>

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

Copyright © 2011 by the authors

Model choice in time series studies of air pollution and mortality

Roger D. Peng

Francesca Dominici

Thomas A. Louis

Corresponding Author: Roger D. Peng (rpeng@jhsph.edu)

Department of Biostatistics

Johns Hopkins Bloomberg School of Public Health

615 N. Wolfe St.

Baltimore, MD 21205

USA

Phone: (410) 955-2468

Fax: (410) 955-0958

Abstract

Multi-city time series studies of particulate matter (PM) 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 United States National Ambient Air Quality Standards (NAAQS) for PM. As a result, methodological issues concerning time series analysis of the relation 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 analyzed 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 commonly used for confounding adjustment. We generate data under several confounding scenarios and systematically compare the performance of the different 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 to the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) database which is comprised of daily time series of several pollutants, weather variables, and mortality counts covering the period 1987–2000 for the largest 100 cities in the United States. When applying these approaches to adjusting for seasonal and long-term trends we find that the NMMAPS estimates for the national average effect of PM_{10} at lag 1 on mortality vary over approximately a two-fold range, with 95% posterior intervals always excluding zero risk.

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

1 Introduction

Numerous time series studies have indicated a positive association between short-term variation of particulate matter and daily mortality counts (see e.g. Pope *et al.*, 1995; Dockery and Pope, 1996; Goldberg *et al.*, 2003; Bell *et al.*, 2004, and references therein). Multi-city studies such as the National Morbidity, Mortality, and Air Pollution Study (Samet *et al.*, 2000c,a), 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 particles, even at levels below current regulatory limits. In the United States, these studies have played an important role in setting standards for acceptable levels of ambient particulate matter. In particular, the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) played a central role in the Environmental Protection Agency's (EPA) development of National Ambient Air Quality Standards (NAAQS) for the six "criteria" pollutants defined by the EPA (Environmental Protection Agency, 1996, 2003).

The critical role of NMMAPS in the development of the NAAQS attracted intense scrutiny from the scientific community and industry groups regarding the statistical models used and the methods employed for adjusting for potential confounding. Confounding occurs when an attribute 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. While collectively strengthening the epidemiologic evidence of the adverse health effects of PM, the proliferation of time series studies employing different approaches to modeling 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 dew point temperature. Daily temperature measurements are readily available for metropolitan areas in the United States 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 dew point) 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 specifically targeted at adjusting for temperature in multi-city 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 $< 10 \mu m$) on total non-accidental mortality is robust to a large class of statistical models used to adjust for potential confounding by temperature and dew point temperature. Building on these findings, in this paper we focus on the problem of controlling for unmeasured confounders, that is, seasonal and long term trends.

Unmeasured confounders are factors that influence mortality and vary with time in a manner 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 fall to early spring and influenza epidemics occur in the same interval but with highly variable timing. The net impact 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/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. TSP 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, dew point 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. A number of 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 only accounts 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. One generally does 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 one should allow for the smooth function of time. This decision is critical because it determines the amount of residual temporal variation in mortality available to estimate the air pollution effect. Over-smoothing the series (thereby under-smoothing the residuals) can leave temporal cycles in the residuals that can produce confounding bias; under-smoothing the series (thereby over-smoothing 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 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 National Morbidity, Mortality, and Air Pollution Study (NMMAPS) database containing daily mortality, pollution, and weather data for 100 U.S. 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 following statistical model:

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

where 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 zero. 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 setup such as (1), in order to estimate β , one must choose how to represent the smooth function f and then decide on the amount of smoothness 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 (1) include natural splines, penalized splines, and smoothing splines. (Other less common choices are loess smoothers or harmonic functions.) The first is fully parametric, while the latter two may be considered more flexible. With natural splines, one constructs a spline basis with knots at fixed locations throughout the range of the data and the choice of knot locations can have a substantial impact 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 datasets such as those used here. Appendix A provides an overview of the different methods used here; a comprehensive treatment can be found in Ruppert *et al.* (2003).

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

- 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).
- GAM-R: The `gam` function in R (from the `mgcv` package) is used with penalized cubic re-

gression 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 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) and Wood (2001).

- 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 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 suggested in Dominici *et al.* (2002). Standard errors are obtained 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 only compare 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, one 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 select a 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) while 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; Touloumi *et al.*, 2004). With each of these approaches, a number of Poisson regression models are fit 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 multi-city 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 we examine here is to use a fixed degrees of freedom, perhaps based on biological knowledge or previous work. For multi-city 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.*, 2000b). One can explore the sensitivity of $\hat{\beta}$ by varying the df 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 degrees of freedom (df) for f .

1. **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 $1, 2, \dots, 20$ df per year of data.
2. **AIC:** Choose the df that minimizes the Akaike Information Criterion (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

$$AIC(df) = -2 \times (\max \text{ log-likelihood}) + 2 df.$$

3. **BIC:** 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; Clyde, 2000). BIC can be written as

$$\text{BIC}(df) = -2 \times (\max \text{ log-likelihood}) + (\log n) df$$

where n is the number of observations.

4. **Minimum residual autocorrelation (PACF):** 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 the df using a test for white noise in the residuals (e.g. Goldberg *et al.*, 2001, and others). While 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 with the df that leads a test for white noise to fail to reject the null hypothesis.
5. **GCV-PM₁₀:** 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 described in Dominici *et al.* (2004).

3 Simulation Study

In this section we describe a simulation study 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 described in Section 2. Our goal is to generate data from confounding scenarios comparable to situations found in real data and evaluate the estimation procedures in each of these scenarios. The definition of the scenarios relies upon the timescales at which confounding occurs and the strength of the concavity between the pollutant series and the seasonal trend. All of the simulations were conducted in R 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:

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

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

In order 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 $a_1, \dots, a_{m_1}$ and a standard linear regression model to the PM_{10} data to obtain estimates of the spline coefficients $b_1, \dots, b_{m_2}$. 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 of the parameters estimated from the Minneapolis/St. Paul data are then treated as the “true” coefficients from which to simulate. The framework in (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 (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; it remains fixed in each of the simulations.

We simulate the following four confounding scenarios:

1. $g(t)$ is smoother than $f(t)$, moderate concavity. 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 $\sigma^2 = \sigma_0^2$.
2. $g(t)$ is smoother than $f(t)$, high concavity. Same as in Scenario 1 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 .
3. $g(t)$ is rougher than $f(t)$, moderate concavity. 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 $\sigma^2 = \sigma_0^2$.
4. $g(t)$ is rougher than $f(t)$, high concavity. Same as in Scenario 3 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 situations attenuate the estimated pollution effect, it may be useful in the future to employ a more elaborate simulation framework to investigate in depth the impact of measurement error.

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

Figure 2 shows boxplots of the 500 estimates of β obtained by using 1 to 20 df per year in the

smooth function of time. The left column shows estimates obtained by using natural splines and right column shows the results of using penalized splines to represent f . The bottom four plots show the estimates obtained under the high concavity scenario (“high cc”). In general, while 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 concavity (top four plots) 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 necessary to represent the g function in (3), then $\hat{\beta}$ is either unbiased (when g is rougher than f) or asymptotically unbiased (g smoother than f). For example, with g rougher than f , we should see very little bias in $\hat{\beta}$ for df larger than or equal to 7 per year. In the nonparametric setting, Rice and Speckman both showed that in order to obtain an estimate of β whose bias converges at the usual parametric rate, one 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 required to obtain an estimate of β with small bias could be less than the amount of smoothing required to obtain a good estimate of f alone (see also Green and Silverman, 1994, ch. 4).

Under high concavity, the differences between using natural splines and penalized splines are greater. For natural splines, the bias drops rapidly between 1 and 4 df per year and is stable afterwards. For penalized splines, the bias drops much more slowly and does not appear to level off until 9 or 10 df 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/rougher) than to the amount of concavity in the data or the basis representation used.

In our comparison of the model selection criteria described in Section 2.2, for each simulated dataset and criterion, we obtain a “best” df , call it $\hat{df}$, that is the value of df associated with the fitted model which minimizes the criterion. The estimate of β chosen by the model selection

criterion for dataset i is $\widehat{\beta}_{df}^{(i)}$. We can then estimate the bias, standard error, and root mean squared error (RMSE) of $\widehat{\beta}_{df}$ from the simulation output for a particular model selection criterion and basis choice. 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 $\widehat{\beta}$ selected by each of the criteria/bases under the different 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 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 concavity, each of the four data-driven methods perform reasonably well with respect to RMSE with BIC always having the largest RMSE. As expected, all of the methods perform worse under high concavity, with BIC having an RMSE 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 obtained via the model selection criteria. Generally, estimates from all of the criteria incur more bias when using penalized splines for the smooth function of time as opposed to natural splines. GCV-PM₁₀ is very nearly unbiased in all of the scenarios. The largest bias (0.159) occurs with penalized splines, under high concavity and when g is smoother than f . AIC has a relatively small bias under the moderate concavity scenarios but tends to incur more bias than GCV-PM₁₀ under the high concavity scenarios (particularly when penalized splines are used). The price for using GCV-PM₁₀ over the other methods appears to be an increase in the standard error of the estimates in some cases.

The PACF criteria performs reasonably well under moderate concavity but has a large bias under high concavity, particularly when using penalized splines. However, the relative increase in bias for the PACF criteria when going from the moderate concavity to the high concavity scenarios is comparable to the other criteria. The BIC criteria performs poorly under all of the scenarios. The larger penalty associated with BIC generally leads to using few degrees of freedom which, from Figure 2, can produce estimates with high bias.

4 NMMAAPS Data Analysis

We apply our methods to the National Morbidity, Mortality, and Air Pollution Study (NMMAAPS) database which is comprised of 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.*, 2000c,a). The data have since been updated to include 10 more cities and 6 more years of data, extending the coverage until the year 2000. The entire database is available via the NMMAAPSdata R package (Peng and Welty, 2004) which can be downloaded from the Internet-based Health and Air Pollution Surveillance System (iHAPSS) website at <http://www.ihapss.jhsph.edu/>.

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

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

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 subscript 1–3). There is also an indicator variable for the day of the week and a separate intercept for each age category (< 65 , $65\text{--}74$, and ≥ 75 years old).

For each city, we choose each of the three fitting procedures (i.e. representations of the smooth function of time) 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 $\hat{df}$, with which we obtain an estimate $\hat{\beta}_{\hat{df}}$ for that city. This process is then repeated for all 100 cities in the database to obtain $\hat{\beta}_{\hat{df}}^{(1)}, \dots, \hat{\beta}_{\hat{df}}^{(100)}$ 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)

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

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 NMMAAPS data. The estimates presented are the national average estimates of the percent increase in mortality for a $10 \mu\text{g}/\text{m}^3$ increase in PM_{10} at lag 1. The results are consistent with what we observed in the simulation studies — AIC and GCV- PM_{10} produce very similar estimates while the PACF estimates are somewhat larger. The estimates obtained by AIC and GCV- PM_{10} are comparable to the estimates reported in previous NMMAAPS analyses (e.g. Dominici *et al.*, 2002, 2003; Peng *et al.*, 2005), although with smaller 95% posterior intervals due to the additional data 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 (a phenomenon common with U.S. data). In particular, for cities where PM_{10} is measured only once every six days, one can only estimate the autocorrelation of the residuals at lag 6. The national average estimates in the second 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.

Figure 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 this figure, rather than minimize one of the model selection criteria and obtain an optimal df in each city, we use a fixed number of degrees of freedom per year for all of the cities. The figure shows the change in the national average estimate as the df is varied. When using natural splines, the estimates appear to stabilize after 9 df per year at around a 0.15% increase in mortality with a $10 \mu\text{g}/\text{m}^3$ increase in PM_{10} at lag 1. The estimates obtained using smoothing splines also appear to stabilize, but at a higher value. The estimates obtained using penalized splines are very close to the smoothing spline estimates up to approximately 12 df per year, after which the penalized spline estimates decrease slightly.

5 Discussion

We have developed a framework for quantifying and characterizing model uncertainty in multi-city time series studies of air pollution and mortality. The complexity of the time series data require the application of sophisticated statistical models capable of estimating relatively small effects. Furthermore, these effects have important policy implications, making a critical evaluation of the diverse modelling approaches 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 associated with each of the different modelling approaches. In addition to the simulation study we have applied all of 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 presented here strengthen recent findings from multi-city 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 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 present 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 impact of concurvity is far greater.

Our results show that both fully parametric and nonparametric methods perform well, with neither preferred. 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), one must use a df that is

not optimal for predicting mortality in order 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 one needs to use a larger df to achieve the same average bias as the corresponding estimate obtained via natural splines (see e.g. Figure 2). Therefore, the automatic use of criteria such as GCV or AIC for selecting the df could be potentially misleading (particularly with high concavity) 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 one must use a df at least as large as that needed to best predict the *pollution series*. They suggested using a procedure such as GCV to estimate this df and then use the bootstrap to minimize an estimate of the mean squared error for $\hat{\beta}$. Our simplified version (GCV-PM₁₀) of their approach performs very well in the simulations and produces estimates of β that are nearly unbiased under all of the scenarios, even under high concavity.

The failure of BIC to produce competitive estimates of β , while dramatic, is not of concern in assessing the relation 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 setup is far from that considered by Schwarz (1978). That is, as $n \rightarrow \infty$, we also have that the dimension of the model $\rightarrow \infty$, 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 concavity, AIC produces estimates of β with relatively small bias. Shibata (1976) demonstrated for autoregressive time series models that AIC has the potential to select larger and 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 $\rightarrow \infty$, AIC can choose the right model as $n \rightarrow \infty$. 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 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 an 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 understimation of standard errors can lead to possibly incorrect inferences about the city-specific coefficients, Daniels *et al.* (2004) showed that in a multi-city context, the underestimation of the city-specific standard errors would have to be severe (or the the number of cities very small) in order 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 $\hat{\beta}$. Figure 2 shows that increasing the 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 the 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 the 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 semi-parametric 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 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 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 datasets 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 of our conclusions from the simulation study are based on assuming a true $\beta = 0$. While 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 zero.

While 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 the U.S. 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 obtained here for the effect of PM₁₀ at lag 1, although the estimates 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 multi-city time series studies are robust to alternative methodologies and data analytic approaches.

6 Acknowledgments

This research was supported in part by NIH/NHLBI grant T32HL07024, NIEHS grant R01ES012054, the NIEHS Center for Urban Environmental Health (P30ES03819). Research described in this article was partially supported by a contract and grant from the Health Effects Institute (HEI), an organization jointly funded by the U.S. Environmental Protection Agency (EPA) and automotive manufacturers. The contents of this article do not necessarily reflect the views and policies of

HEI, nor do they necessarily reflect the views and policies of EPA, nor motor vehicle or engine manufacturers.

References

- Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, 267–281.
- Bell, M. L., Samet, J. M. and Dominici, F. (2004) Time-series studies of particulate matter. *Annual Review of Public Health*, **25**, 247–280.
- Berger, J. O., Ghosh, J. K. and Mukhopadhyay, N. (2003) Approximations and consistency of Bayes factors as model dimension grows. *Journal of Statistical Planning and Inference*, **112**, 241–258.
- Brockwell, P. J. and Davis, R. A. (2002) *Introduction to Time Series and Forecasting*. Springer, 2nd edn.
- Buja, A., Hastie, T. and Tibshirani, R. (1989) Linear smoothers and additive models. *Annals of Statistics*, **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. *Canadian Journal of Public Health*, **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*, 85–89. Health Effects Institute, Cambridge MA.
- 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. *American Journal of Epidemiology*, **155**, 80–87.

- Daniels, M., Dominici, F., Samet, J. M. and Zeger, S. L. (2000) Estimating PM₁₀-mortality dose-response curves and threshold levels: An analysis of daily time-series for the 20 largest US cities. *American Journal of Epidemiology*, **152**, 397–412.
- 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), 123–147. 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*, 9–24. The Health Effects Institute, Cambridge MA.
- Dominici, F., McDermott, A. and Hastie, T. (2004) Improved semiparametric time series models of air pollution and mortality. *Journal of the American Statistical Association*, **99**, 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. *American Journal of Epidemiology*, **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). *Journal of the Royal Statistical Society, Series A*, **163**, 263–302.
- Donnell, D. J., Buja, A. and Stuetzle, W. (1994) Analysis of additive dependencies and concurvities using smallest additive principal components. *Annals of Statistics*, **22**, 1635–1668.
- Eilers, P. H. C. and Marx, B. D. (1996) Flexible smoothing using *B*-splines and penalized likelihood (with discussion). *Statistical Science*, **11**, 89–121.
- Environmental Protection Agency (1996) *Air Quality Criteria for Particulate Matter*. EPA/600/P-95/001aF. Office of Research and Development, Washington DC.

- (2003) *Air Quality Criteria for Particulate Matter (Fourth External Review Draft)*. EPA/600/P-99/002aD and bD. Office of Research and Development, National Center for Environmental Assessment, Research Triangle Park, NC.
- Everson, P. J. and Morris, C. N. (2000a) Inference for multivariate normal hierarchical models. *Journal of the Royal Statistical Society, Series B*, **62**, 399–412.
- (2000b) Simulation from Wishart distributions with eigenvalue constraints. *Journal of Computational and Graphical Statistics*, **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. *American Journal of Epidemiology*, **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 Health*, **18**, 269–303.
- Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall, London UK.
- Gu, C. (2002) *Smoothing Spline ANOVA Models*. Springer, New York.
- Hansen, M. and Yu, B. (2003) Minimum description length model selection criteria for generalized linear models. In *Science and Statistics: A Festschrift for Terry Speed* (ed. D. Goldstein), 40, 145–164. IMS Lecture Notes.
- 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*. Health Effects Institute, Boston MA.
- 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**, 521–513.
- Kelsall, J. E., Samet, J. M., Zeger, S. L. and Xu, J. (1997) Air pollution and mortality in Philadelphia, 1974–1988. *American Journal of Epidemiology*, **146**, 750–762.
- Marx, B. D. and Eilers, P. H. C. (1998) Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis*, **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. *American Journal of Epidemiology*, **161**, 585–594.
- Peng, R. D. and Welty, L. J. (2004) The NMMAData package. *R News*, **4**, 10–14. URL <http://CRAN.R-project.org/doc/Rnews/>.
- Pope, C. A., Dockery, D. and Schwartz, J. (1995) Review of epidemiological evidence of health effects of particulate air pollution. *Inhalation Toxicology*, **47**, 1–18.
- R Development Core Team (2003) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org>. ISBN 3-900051-00-3.
- 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. *Statistics and Probability Letters*, **4**, 203–208.
- Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press.
- Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. Cambridge University Press.

- Samet, J. M., Zeger, S. L., Dominici, F., Currier, I., 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*. Health Effects Institute, Cambridge MA.
- (2000b) *The National Morbidity, Mortality, and Air Pollution Study, Part II: Morbidity and Mortality from Air Pollution in the United States*. Health Effects Institute, Cambridge MA.
- Samet, J. M., Zeger, S. L., Dominici, F., Dockery, D. and Schwartz, J. (2000c) *The National Morbidity, Mortality, and Air Pollution Study, Part I: Methods and Methodological Issues*. Health Effects Institute, Cambridge MA.
- Samet, J. M., Zeger, S. L., Kelsall, J., Xu, J. and Kalkstein, L. (1998) Does weather confound or modify the association of particulate air pollution with mortality? *Environmental Research, Section 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. *Environmental Health Perspectives*, **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. *Occupational and Environmental Medicine*, **60**, 977–982.
- Schwartz, J. (1994a) Nonparametric smoothing in the analysis of air pollution and respiratory illness. *Canadian Journal of Statistics*, **22**, 471–488.
- (1994b) Total suspended particulate matter and daily mortality in Cincinnati, Ohio. *Environmental Health Perspectives*, **102**, 186–189.
- (2000) The distributed lag between air pollution and daily deaths. *Epidemiology*, **11**, 320–326.

- (2004) Is the association of airborne particles with daily deaths confounded by gaseous air pollutants? An approach to control by matching. *Environmental Health Perspectives*, **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*, 25–58. Health Effects Institute, Cambridge MA.
- Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **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. *Journal of the Royal Statistical Society, Series B*, **50**, 413–436.
- Stone, M. (1979) Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society, Series 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) Flexible distributed lag models: Are the acute effects of PM₁₀ on mortality the result of inadequate control for weather and season? *American Journal of Epidemiology*, to appear.
- Wood, S. N. (2000) Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B*, **62**, 413–428.
- (2001) mgcv: GAMs and generalized ridge regression for R. *R News*, **1/2**, 20–25.

A Details of Representations for f

Natural splines are piecewise cubic polynomials defined on a grid of knot locations spanning the range of the data. The function itself, as well as its second derivative, are continuous on the entire range of the data and the function is restricted to be linear beyond the endpoints. The smoothness of a natural spline fit is controlled by the number of knots used. Fewer knots represent smoother fits while 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 a number of ways and we use the general definition, $\hat{\boldsymbol{\eta}}^T \mathbf{B}(x)$, where

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

$\mathbf{B}(x)$ 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 generally considered necessary) and then shrinks the coefficients to reduce the effective degrees of freedom and increase 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, that is, 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 dataset. Since smoothing splines can be considered a special case of penalized splines (Ruppert *et al.*, 2003), we expect that results obtained 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 e.g. 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 will take more simplified version of (1), focusing on the estimation of β and the smooth function of time f . Using matrix notation, we can rewrite (1) as

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

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

Given one of the spline bases described in Section 2.1, we can rewrite (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 (IRLS) to fit model (5) 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 IRLS algorithm. Let X^* be the complete design matrix, i.e. $X^* = [X \mid B]$. Using a GLM procedure with natural cubic splines, we can estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ simultaneously as

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

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

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

and the estimate of β is

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

Rice (1986) and Speckman (1988) both showed that while the variance of $\hat{\beta}_{ps}$ converges at the standard parametric rate for $n \rightarrow \infty$, the bias converges to zero 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 $\hat{\beta}_{ns}$ and $\hat{\beta}_{ps}$ using various model selection criteria was illustrated in Sections 3 and 4.

Speckman described an alternative estimator for β for which the bias and variance both converge at the usual parametric rate. For S symmetric, the modified estimator is

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

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

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

which is the estimate one 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 β using $\hat{\beta}_{ps}^*$ is equivalent to estimating β with $\hat{\beta}_{ps}$ but with the modified smoother matrix

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

Buja *et al.* showed that $\text{tr}(\tilde{S}) \geq \text{tr}(S)$ and hence, in order to obtain an estimate of β for which the bias and variance converge at the parametric rate, one must implicitly use an undersmoothed

estimate of $\mathbf{f}$. If Λ is the diagonal matrix of eigenvalues $\lambda_1, \dots, \lambda_n$ for the smoother matrix S , a simple calculation reveals that

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

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

B Tables and Figures

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

Table 1: Simulation scenarios: $\sigma_0^2 = 186.7$ for scenarios where $g(t)$ is smoother than $f(t)$ and $\sigma_0^2 = 182.2$ for scenarios where $g(t)$ is rougher than $f(t)$.

Figure 1: Example of simulated mortality and PM₁₀ data. The negative values in the PM₁₀ 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 concavity.

	Basis	Natural splines				Penalized splines			
	Concurvity $g(t)$	Moderate		High		Moderate		High	
		smoother	rougher	sm	r	sm	r	sm	r
Bias ($\times 1000$)	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
Std. Error ($\times 1000$)	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 ($\times 1000$)	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

Table 2: Average bias, standard error, and root mean squared error (RMSE) of $\hat{\beta}$ (all $\times 1000$) from 250 simulations. Each column represents a scenario determined by the basis used for fitting (natural splines/penalized splines), the concurvity in the simulated data, and the relationship between $g(t)$ and $f(t)$, i.e. $g(t)$ smoother (sm) or rougher (r) than $f(t)$.

	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)

Table 3: National average estimates and 95% posterior intervals of the percent increase in mortality with a 10 $\mu\text{g}/\text{m}^3$ increase in PM₁₀ at lag 1 using different model selection criteria and representations of the smooth function of time, $f(t)$.

Figure 2: Sensitivity analysis of $\hat{\beta}$. The boxplots show the distribution of $\hat{\beta}$ over 250 simulations using 1–20 df per year in the smooth function of time f (the true $\beta = 0$). The methods used were GLM with natural cubic splines (GLM) and GAM with penalized splines (GAM-R).

Figure 3: Sensitivity analysis of the national average estimate of the percent increase in mortality for a $10 \mu\text{g}/\text{m}^3$ increase in PM_{10} at lag 1. City-specific estimates were obtained from 100 U.S. cities using data for the years 1987–2000 and the estimates were combined using a hierarchical normal model. The three fitting methods used are GLM with natural cubic splines (GLM-NS), GAM with penalized splines (GAM-R), and GAM with smoothing splines (GAM-S). The shaded region shows 95% posterior intervals for the estimates obtained using GLM-NS.