Time-varying coefficient models for the analysis of air pollution and health outcome data

Duncan Lee and Gavin Shaddick

Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY,

United Kingdom.

SUMMARY. In this paper a time-varying coefficient model is developed to examine the relationship between adverse health and short-term (acute) exposure to air pollution. This model allows the relative risk to evolve over time, which may be due to an interaction with temperature, or from a change in the composition of pollutants, such as particulate matter, over time. The model produces a smooth estimate of these time-varying effects, which are not constrained to follow a fixed parametric form set by the investigator. Instead, the shape is estimated from the data using penalised natural cubic splines. Poisson regression models, using both quasi-likelihood and Bayesian techniques, are developed, with estimation performed using an iteratively re-weighted least squares procedure and Markov chain monte carlo (MCMC) simulation respectively. The efficacy of the methods to estimate different types of time-varying effect are assessed via a simulation study, and the models are then applied to data from four cities which were part of the National Morbidity, Mortality, and Air Pollution Study (NMMAPS).

KEY WORDS: Air pollution; Bayesian hierarchical models; Epidemiology; Penalised splines; Time-varying coefficient models.

1. Introduction

The potential association between exposure to air pollution and adverse health events has been a major issue in public health for over fifty years. Numerous studies have shown a link between a component of air pollution and measures of public health, the majority of which

email: d.p.lee@maths.bath.ac.uk

have estimated the effects associated with short-term exposure to air pollution. These studies are typically based on daily data from a specific region and time period, and analysis is carried out using time series regression methods, such as Poisson linear (McCullagh and Nelder (1989)) and additive (Hastie and Tibshirani (1990)) models. These models make a number of, possibly untenable, assumptions about the underlying mechanism that generates the daily health data, and relaxing these assumptions is an active research topic. Such assumptions include: the same mean and variance for the health data; independence of daily health counts; a constant effect of air pollution; and modelling seasonal variation using fixed parametric functions. Advances in statistical methods have allowed researchers to relax these assumptions and apply a wider class of regression models to these data.

Recent research has investigated the shape of the air pollution effects, with particular emphasis on comparing a constant effect with a dose response relationship (Daniels et al. (2000)). Such a relationship allows the effects of air pollution to depend on the pollution level, but comparatively little research has investigated the possibility of temporal variation in the effects of air pollution. This variation may be seasonal, which could result from an interaction with temperature or with another pollutant exhibiting a seasonal pattern. Alternatively, it may exhibit a long-term trend, which could result from such interactions, a change in the composition of individual pollutants over a number of years, or from a change in the size and structure of the population at risk. In this paper we propose a time-varying coefficient model, that estimates the temporal variation in the effects of pollution using a penalised natural cubic spline. Bayesian and quasi-likelihood implementations of this model are presented, with particular interest in the differences between the two estimates and associated confidence and credible intervals.

The remainder of this paper is organised as follows. Section 2 describes the models previously used to estimate the association between health effects and short-term exposure to air pollution, and proposes a time-varying coefficient model for such analysis. Section 3 describes a Bayesian analysis with inference based on Markov Chain Monte Carlo (MCMC) simulation, and a likelihood based alternative, which uses an iteratively re-weighted least squares procedure. In section 4 a simulation study is carried out to determine if the models proposed in this paper can accurately estimate different shaped time-varying effects. Section 5 applies these models to real data from four U.S. cities over a five year period (1993-1997). Finally section 6 gives a concluding discussion and suggests some future extensions.

2. Modelling the effects of air pollution on public health

The adverse health effects associated with acute exposure to air pollution are typically estimated from daily ecological data that relate to a specific region for n consecutive days. The data comprise counts of adverse health events (such as mortality) $\mathbf{y} = (y_1, \ldots, y_n)^T_{n\times 1}$, levels of air pollution $\mathbf{x} = (x_1, \ldots, x_n)^T_{n\times 1}$, and a matrix of covariates $Z = (\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T)^T_{n\times q}$. The covariates model confounding factors, such as long-term trends, seasonal variation and serial correlation in the health data, and typically include smooth functions of calendar time and temperature. Regression models for these data are based on Poisson generalised linear or additive models, and the overall form depends on the methods chosen to estimate the smooth functions and the form of the air pollution-health relationship. In this paper we take a parametric approach to modelling confounders using natural cubic splines, (allowing straightforward estimation in a Bayesian setting) so a common general form can be expressed as

$$
\begin{aligned}
y_t &\sim \text{Poisson}(\mu_t) \quad \text{for} \;\; t = 1, \ldots, n, \\
\log(\mu_t) &= x_t \beta_t + \mathbf{z}_t^{\mathrm{T}} \boldsymbol{\delta}, \\
\beta_t &= f(t; \boldsymbol{\alpha}).
\end{aligned}
\tag{1}
$$

The effect of air pollution on day t is represented by β_t, and the evolution over time is modelled by a function f with parameter vector $\boldsymbol{\alpha}$. The covariates include basis functions for natural cubic splines, and their effects are controlled by the $q \times 1$ parameter vector $\boldsymbol{\delta}$, the first element of which is an intercept term.

2.1 *Specific forms of the air pollution and health relationship*

A number of previously used forms for f are discussed below, together with the time-varying coefficient model proposed here.

(i) $\beta_t = \alpha_1$, for a constant effect of air pollution.

(ii) $\beta_t = 1$ and x_t is replaced with $f(x_t; \lambda)$, for a dose response relationship.

(iii) $\beta_t = \alpha_0 + \alpha_1 \sin(2\pi t/365) + \alpha_2 \cos(2\pi t/365)$, for a smooth seasonal time-varying effect of air pollution.

(iv) $\beta_t = \beta_{t-1} + \omega_t$ and $\omega_t \sim N(0, \alpha_1)$, for a time-varying effect of air pollution modelled as a first order random walk.

(v) $\beta_t = f(t; \boldsymbol{\alpha})$, where f is an arbitrary function that estimates a smooth time-varying effect of air pollution.

Non time-varying effects (i)-(ii)

The majority of researchers assume the effect of air pollution is either constant (see for example Mar et al. (2000) and Moolgavkar (2000)), or varies as a non-linear function of the quantity of air pollution, a so-called dose-response relationship (see for example Schwartz (1994) and Daniels et al. (2000)).

Time-varying effects (iii)-(iv)

Other extensions, as considered here, have allowed the relationship between air pollution and health to change over time. The only known analyses are those by Moolgavkar et al. (1995) and Peng et al. (2005) (model iii) who proposed seasonal models, and those by Chiogna and Gaetan (2002) (model iv) and Lee and Shaddick (2005), who adopt an autoregressive approach. The seasonal models fix the parametric form of the time-varying effects *a-priori*, while the autoregressive models do not force the evolution of the time-varying effects to be smooth.

2.2 *Time varying coefficient models* (v)

A time varying coefficient model (TVCM) is a special case of a varying coefficient model (Hastie and Tibshirani (1993)), for which the effect modifier is time. The model proposed here is that of equation (1), with the vector of air pollution effects, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)_{n\times 1}$, modelled as an arbitrary smooth function $f(t; \boldsymbol{\alpha})$. The advantage of this approach, over those discussed in the previous section, is that $\boldsymbol{\beta}$ is completely smooth, with its shape determined from the data and not from the parametric form specified by the investigator. We estimate the smooth function with a regression spline because it is fully parametric, making implementation within a Bayesian framework relatively straightforward. We use a natural cubic spline because it is visually smooth, and the shape beyond the two end knots is constrained to be linear, precluding any erratic tail behaviour. An alternative, not discussed here, is to estimate $f(t; \boldsymbol{\alpha})$ with a non-parametric function such as a smoothing spline, with estimation based on the methods discussed by Lin and Zhang (1999). A regression spline comprises a linear combination of p basis functions, $f(t; \boldsymbol{\alpha}) = \mathbf{B}_t^T \boldsymbol{\alpha}$, where $\mathbf{B}_t = (B_1(t), \ldots, B_p(t))_{p\times 1}$ is a vector of known basis functions evaluated at day t, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)_{p\times 1}$ is a vector of regression parameters. In this paper we use a B-spline basis (Eilers and Marx (1996)), because it is numerically stable and implementation within Bayesian and likelihood frameworks is straightforward.

We use a penalised approach to estimation because it allows the smoothing to be controlled by a single parameter, rather than by specifying the size and location of a set of knots. This approach uses an overly large set of knots and penalised excess curvature in the estimate via a penalty term. The form of this penalty depends on the set of basis functions, and for likelihood estimation we use the suggestion of Eilers and Marx (1996), who penalise kth order differences between the coefficients of the spline. For example, second order differences are given by $\sum_{j=3}^{p}(\alpha_j - 2\alpha_{j-1} + \alpha_{j-2})^2 = \boldsymbol{\alpha}^T D\boldsymbol{\alpha}$, where

$$D = \begin{pmatrix} 1 & -2 & 1 & & & & & & \\ -2 & 5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & \\ & 1 & -4 & 6 & -4 & 1 & & & \\ & & \vdots & \vdots & \vdots & \vdots & \vdots & & \\ & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & 1 & -4 & 6 & -4 & 1 \\ & & & & & 1 & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{pmatrix}. \quad (2)$$

The penalty term is multiplied by a smoothing parameter λ, with larger values leading to the fitted curve being smoother. Penalised B-splines have been adapted to a Bayesian setting by Lang and Brezger (2004), who replaced the difference penalty of order k with a kth order random walk prior, its stochastic analogue. For example, a second order random walk prior is given by $\alpha_j|\alpha_{j-1}, \alpha_{j-2}, \tau^2 \sim \mathrm{N}(2\alpha_{j-1} - \alpha_{j-2}\ ,\ \tau^2)$ for $j = 3, \ldots, p$, with non-informative priors for (α_1, α_2). The smoothing parameter is τ^2, which controls the size of the differences between the parameters $\boldsymbol{\alpha}$, with larger values leading to less smoothing.

3. Statistical inference

This section describes Bayesian and likelihood approaches to estimation.

3.1 *Bayesian analysis*

The Bayesian model is that of equation (1), with the vector of air pollution effects modelled by a penalised spline, $\beta_t = \mathbf{B}_t^T \boldsymbol{\alpha}$, as suggested in section 2. The penalty takes the form of a second order random walk prior, $\alpha_j|\alpha_{j-1}, \alpha_{j-2}, \tau^2 \sim \mathrm{N}(2\alpha_{j-1}-\alpha_{j-2}\ ,\ \tau^2)$, with non-informative priors for (α_1, α_2). The vector of covariate parameters has a multivariate Gaussian prior, $\boldsymbol{\delta} \sim \mathrm{N}(\mathbf{m}, \Sigma)$, and the smoothing parameter τ^2 has a conjugate Inverse-Gamma prior, $\tau^2 \sim$ Inverse-Gamma(e, f). The hyperparameters $(\mathbf{m}_{q\times 1}, \Sigma_{q\times q}, e, f)$ are known and chosen to make the priors non-informative. Inference is performed via MCMC simulation using a hybrid Metropolis-within Gibbs approach that updates the parameters in blocks $[\boldsymbol{\delta} = (\delta_1, \ldots, \delta_q), \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p), \tau^2]$. The joint posterior distribution is given by

$$p(\boldsymbol{\alpha}, \boldsymbol{\delta}, \tau^2|\mathbf{y}) \propto \prod_{t=1}^{n} \mathrm{Poisson}(y_t|\boldsymbol{\alpha}, \boldsymbol{\delta}) \prod_{j=3}^{p} \mathrm{N}(\alpha_j|2\alpha_{j-1}-\alpha_{j-2}, \tau^2)\mathrm{N}(\boldsymbol{\delta}|\mathbf{m}, \Sigma)\text{Inverse-Gamma}(\tau^2|e, f).$$

Simulation for this type of regression problem has been developed by Fahrmeir and Lang (2001), and the simulation algorithm adopted here is based on their work. Details are given in web appendix A.

3.2 *Likelihood based analysis*

The likelihood based analysis is based on equation (1), with the vector of air pollution effects modelled by a penalised spline, $\beta_t = \mathbf{B}_t^T \boldsymbol{\alpha}$, as suggested in section 2. The spline parameters $\boldsymbol{\alpha}$ are subject to the second order difference penalty described in section 2, and $(\boldsymbol{\alpha}, \boldsymbol{\delta})$ can be estimated using the penalised likelihood approach suggested by Marx and Eilers (1998). Further details are given in web appendix B.

4. Simulation study

In this section we describe a simulation study, to assess the effectiveness of the time-varying coefficient model described in section two. Specifically, we simulate four sets of mortality data with different types of time-varying effect: (i) constant; (ii) seasonal with a period of a year; (iii) a quadratic trend; (iv) a smooth cubic spline with 6 degrees of freedom. Each simulated data set comprises daily counts of mortality for a three year period, which are generated from a Poisson regression model. The vector of Poisson mean values depends on air pollution data from Detroit and a set of covariates, the latter of which include an intercept term, cyclical components with periods of a whole, half and a quarter of a year, and a natural cubic spline of temperature (also from Detroit) with 3 degrees of freedom. The time-varying effects of air pollution are chosen to be a similar size to those found in current studies, with a relative risk around 1% for an increase of $10\mu g/m^3$.

4.1 *Results*

The time-varying effects from the simulated mortality data are estimated using the Bayesian and likelihood penalised spline models described in section three. We apply our models to two different sets of covariates: (i) the exact set of covariates used to simulate the data; (ii) a set of covariates chosen by model building criteria and residual based methods. We use the first to ensure our models accurately estimate different types of effect, while the second represents the standard situation where the set of confounders are unknown. The standard approach to controlling confounding uses smooth functions of calendar time and temperature, and for these data we use natural cubic splines. We use deviance information criteria (DIC, Spiegelalter et al. (2002)) to select the degrees of freedom, and end up with fifteen (five per year) for calendar time and two for temperature. Additionally, we also include an intercept term. The Bayesian estimates are based on the posterior median from 130,000 samples, which are burnt in for 30,000 iterations and then thinned by five to give

20,000 final samples. The smoothing parameter for the likelihood estimates is chosen by generalised cross validation (GCV), although Akaike's information criterion (AIC) gives similar results. Figures 1 and 2 show the actual time varying effects (on the relative risk scale for an increase of $10\mu g/m^3$), together with those estimated from the Bayesian and likelihood methods using the exact covariates and those chosen by model building criteria.

[Figure 1 about here.]

[Figure 2 about here.]

The Bayesian and likelihood methods estimate the underlying time-varying effects well, showing the correct overall shape for each set of data. The use of the exact covariates improves the estimates of a constant effect (panel (i)), but has little impact on the remaining three time-varying estimates. Although all the underlying shapes are well estimated, the seasonal effect is estimated with the least accuracy, suggesting that the proposed models perform worse if the underlying temporal variation has greater curvature. The Bayesian (dotted line) and likelihood (dashed line) estimates are very similar for most data sets, but when the estimates differ, neither is preferable (see for example the quadratic and seasonal effects using the exact set of covariates). When the covariates are chosen by model building criteria, the estimates from the constant and spline models are similar, suggesting that in this standard situation, the models may struggle to distinguish between a constant effect and a slowly evolving long-term trend.

5. Application

The Bayesian and likelihood models described in sections two and three are illustrated by analysing the relationship between particulate matter and mortality in four U.S. cities.

5.1 *Data*

The models presented here are applied to daily mortality and particulate matter data, which were first analysed in the National Morbidity, Mortality and Air Pollution Study (NMMAPS, Samet et al. (2000)). The four cities are Cleveland, Detroit, Minneapolis and Pittsburgh, and we obtain the data from the R package 'NMMAPSdata' (Peng and Welty (2004)). We use the data on these cities from 1^{st} January 1993 until the 31^{st} December 1997, because daily PM_{10} and weather data have the fewest missing values of all the 108 NMMAPS cities over the fourteen years of available data (1987 - 2000). The response data $\mathbf{y}$ comprise daily counts of total non-accidental mortality for all age groups, and the PM_{10} data are daily averages across a number of monitors in each city. The weather data include daily mean temperature and mean dewpoint temperature.

5.2 *Statistical models*

We compare our Bayesian and likelihood implementations of the penalised spline model against some of the simpler alternatives discussed in section two. In particular, we apply five models to each data set.

Model 1 - Bayesian penalised spline model, $\beta_t = \mathbf{B}_t^T \boldsymbol{\alpha}$, with estimation carried out as described in section 3.1.

Model 2 - Likelihood penalised spline model, $\beta_t = \mathbf{B}_t^T \boldsymbol{\alpha}$, with estimation carried out as described in section 3.2.

Model 3 - Constant effect of pollution $\beta_t = \alpha_1$.

Model 4 - Seasonal effect of pollution $\beta_t = \alpha_0 + \alpha_1 \sin(2\pi t/365) + \alpha_2 \cos(2\pi t/365)$.

Model 5 - Cubic effect of pollution $\beta_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3$.

Analyses for models 3-5 are carried out using Bayesian methods based on MCMC simulation. The covariates are the same for all five models, and are given by

$$\begin{aligned}
\mathbf{z}_t^T \boldsymbol{\delta} &= \delta_1 + f(\mathrm{t}; 35) + f(\mathrm{temperature}_0; 6) + f(\mathrm{temperature}_{1-3}; 6) \\
&\quad + f(\mathrm{dewpoint}_0; 3) + f(\mathrm{dewpoint}_{1-3}; 3) + \mathrm{DOW}\boldsymbol{\delta}^*,
\end{aligned}$$

in which $f(var; df)$ denotes a natural cubic spline of the variable 'var' with 'df' degrees of freedom. The first term is an intercept (δ_1), and the variables 't', 'temperature' and 'dewpoint' denote calendar time, mean temperature, and mean dewpoint temperature respectively. The subscripted numbers represent the 'lag' used, so that '$\mathrm{temperature}_0$' is the same days temperature and '$\mathrm{dewpoint}_{1-3}$' is dewpoint temperature averaged over lags one to three. Finally 'DOW' denotes six indicator variables for day of the week, with Monday taken to be the default. The collection of parameters that comprise the splines, the intercept and the day of the week effects are collectively denoted by $\boldsymbol{\delta}$. The next subsection describes our model building process which led to the choice of covariates, and prior distributions.

5.2.1 Model building and justification Model building is carried out using Bayesian generalised linear models and MCMC simulation. We start by modelling the abundant seasonal variation and serial correlation in the daily mortality counts using natural cubic splines of calendar time, temperature and dewpoint temperature. Splines of calendar time, with a range of degrees of freedom, were fitted to the data, and we use 35 degrees of freedom (seven per year) as it produces the lowest DIC and little structure in the Bayesian residuals. After calendar time, we then investigate different lags and moving averages of temperature and dewpoint temperature, and use the current day and a one to three day moving average because they are significant (posterior estimate was significantly different from zero), and have low DIC for all four cities. We include these variables as natural cubic splines with low degrees of freedom, because temperature variables typically show a U-shaped relationships with mortality (see for example Schwartz (1994)). Finally, we add indicator variables for day of the week to the model as they have a significant effect on daily mortality. We assign a non-informative multivariate Gaussian prior for the vector of covariate parameters, with a diagonal variance matrix and mean vector based on data from earlier years (1990-1992).

After modelling the confounding factors, we add air pollution to the regression model. We include a one day lag because it has the lowest DIC for three of the four cities and the most significant posterior estimates. The time-varying effects of pollution are based on 60 basis functions (12 per year), which allows any non-linearity in the time-varying effect to be captured. Sensitivity analysis shows that the results do not change if we use up to 100 basis functions. Sensitivity analysis also shows that the posterior for τ^2 depends on the choice of Inverse-Gamma(ϵ, ϵ) prior, and we use a non-informative prior on the standard deviation scale as suggested by Gelman (2006).

5.3 *Results*

The Bayesian models are implemented using two parallel chains of 165,000 samples, which are 'burnt in' for 40,000 iterations and thinned by 5, resulting in 25,000 samples from each chain. Convergence is monitored by the methods of Gelman et al. (2003), and the starting values are generated from overdispersed versions of the priors. Likelihood based estimation of the smoothing parameter λ is carried out by GCV, although AIC gives similar answers. Plots of the respiratory mortality data and the median fitted values from the Bayesian spline model (Model 1) are shown for the four data sets in Figure 3. The fitted values from Models 2-5 are similar and are not shown. All models capture the underlying seasonal trend in mortality well, and the residuals show little structure or correlation (based on residual plots and the autocorrelation (ACF) and partial autocorrelation (PACF) functions not shown).

[Figure 3 about here.]

5.3.1 Time-varying effects of PM_{10} The time-varying effects of PM_{10} at lag one are shown in Figures 4 (Cleveland and Detroit) and 5 (Minneapolis and Pittsburgh). The left columns correspond to Cleveland and Minneapolis, while the right columns represent Detroit and Pittsburgh. In each column, panels (i) and (ii) show the time-varying effects and corresponding confidence and credible intervals for the Bayesian and likelihood spline models (1 and 2), while panel (iii) give the constant (dashed), seasonal (dotted) and trend models (3,4,5).

[Figure 4 about here.]

[Figure 5 about here.]

The overall relative risks, as measured by the constant model (Cleveland - 1.0049, Detroit - 1.0046, Minneapolis - 1.0052, Pittsburgh - 1.0045), are similar to those estimated from previous analyses of these data (see for example Samet et al. (2000)). The estimated time-varying effects from both spline models (Models 1 and 2) are not consistent over the four cities, although they all exhibit long-term trends, both increasing (Detroit) and decreasing (Pittsburgh) over time. In particular, these models show no seasonal pattern for any of the four cities, which is in contrast to the work of Peng et al. (2005), who used a model which forced the effects to adopt a sinusoidal shape with a period of one year. The seasonal model we use is similar to that proposed by Peng et al. (2005), and the estimated seasonal effects appear to be spurious because they are not corroborated by either spline model. The Bayesian and likelihood spline models produce estimates that are similar in shape, with the Bayesian estimate showing slightly more curvature over the five year period. The biggest difference can be seen in Minneapolis, where the Bayesian model estimates a peak in the effects of PM_{10} in the winter of 1993/94, which contrasts with the steady increase estimated by the likelihood model. The Bayesian 95% pointwise credible intervals for the spline models are generally wider than the corresponding likelihood interval, but the differences are not large. The cubic model (Model 5) estimates are very similar to the spline models, suggesting that if the time-varying effects exhibit a long-term trend, the simple cubic model performs as well as the penalised spline approach.

[Table 1 about here.]

The posterior distributions of the Bayesian smoothing parameters are summarised in Table 1. The majority of the posterior mass is close to zero, which represents close to maximal smoothing. In the limit of maximal smoothing, the estimated time-varying effect would behave like a low order polynomial, which can be seen by its similarity to the estimated effect from the cubic model. For the likelihood model, the smoothing parameter was estimated by generalised cross validation (GCV), and the estimates (not given) result in near-maximal smoothing.

6. Discussion

The regression models proposed in this paper allow the effects of air pollution to vary smoothly over time without restricting their temporal shape. These effects may show a seasonal pattern, which could result from an interaction with temperature, or exhibit a long-term trend, which could be caused by a change in the composition of individual pollutants over a number of years. The time-varying effects are modelled using a penalised natural cubic spline, with implementation in both Bayesian and likelihood settings. The adequacy of the models are illustrated by a simulation study, and are subsequently applied to data from four U.S. cities obtained from the National Morbidity, Mortality and Air Pollution Study.

The results from the simulation study show that the Bayesian and likelihood penalised spline models estimate a variety of time-varying effects closely, picking out constant effects, long-term trends, and cyclical variation. The estimates are more accurate when the exact set of covariates are used, but retain the correct overall shape in the standard setting where the set of covariates are unknown. The Bayesian and likelihood estimates are very similar, and are more accurate if the time-varying effects exhibit less curvature. However, if the set of confounding variables are unknown, a constant effect may be indistinguishable from a slowly increasing or decreasing trend.

In the four U.S. cities studied, the overall increase in mortality, estimated as a relative risk for an increase of $10\mu g/m^3$ in PM_{10}, was around 0.5%, which is similar to previous analyses of these data. The Bayesian and likelihood time-varying estimates are similar, showing the same pattern of increasing or decreasing long-term trends across the four cities. However, in light of the simulation study, a constant effect that does not vary over time cannot be ruled out. The Bayesian estimates exhibit slightly more curvature than their likelihood counterparts for each city, which is probably a result of the estimation techniques. The likelihood approach calculates the likelihood for a range of values of the smoothing parameter, and estimates λ by optimising a data driven criterion. In contrast, the Bayesian approach averages over the prior for τ^2, which incorporates the possibility of no smoothing, and thus leads to a less smooth estimate. The cubic model produces similar results to our penalised spline approach, which suggests if a constant effect or long-term trend is present, the cubic model is equally good. However, the advantage of our penalised spline model is its flexibility (as shown in section 4) to detect a variety of temporally varying effects, which is beyond the range of the constant, cubic and seasonal models, which are restricted by their fixed parametric form.

The Bayesian 95% credible intervals are generally wider than the corresponding likelihood confidence intervals, which probably results from the estimation of the respective smoothing parameters (λ, τ^2). The Bayesian model correctly allows the uncertainty associated with the smoothing parameter to be incorporated into the model, leading to more realistic estimates of the variability in the effects of air pollution. In contrast the likelihood confidence intervals are likely to be too narrow, because the smoothing parameter is assumed to be estimated without error.

At present, the models are applied separately to data from four individual cities. Future development could extend the methodology to deal with data from multiple cities simultaneously, with the aim of estimating regional and national time-varying effects. This could be achieved within a Bayesian hierarchical model, although the computational burden may restrict the choice of spatial model. A further avenue of research would be to estimate the time-varying effects using a non-parametric smooth function such as a LOESS smoother or smoothing spline, to determine if their increased flexibility compared with a regression spline produced different estimates.

7. Supplementary materials

All supplementary materials for this article can be downloaded as a single document from the Biometrics website at http://www.tibs.org/biometrics. This document describes the MCMC (Web appendix 1) and penalised least squares (Web appendix 2) algorithms, that were used to implement the Bayesian and likelihood analyses.

Acknowledgements

We would like to thank the referees, the associate editor and the editor whose constructive comments greatly improved the focus and content of this work.

References

Chiogna, M. and Gaetan, C. (2002). Dynamic generalized linear models with applications to environmental epidemiology. *Applied Statistics* **51**, 453–468.

Daniels, M., Dominici, F., Samet, J. and Zeger, S. (2000). Estimating Particulate Matter-Mortality Dose-Response Curves and Threshold Levels: An Analysis of Daily Time-Series for the 20 Largest US Cities. *American Journal of Epidemiology* **152**, 397–406.

Eilers, P. and Marx, B. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science* **11**, 89–1218.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov ranodm field priors. *Applied Statistics* **50**, 201–220.

Gelman, A. (2006). Prior distirbutions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515–533.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2003). *Bayesian Data Analysis.* Chapman and Hall, 2nd edition.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, 1st edition.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient Models. *Journal of the Royal Statistical Society Series B* **55**, 757–796.

Lang, S. and Brezger, A. (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.

Lee, D. and Shaddick, G. (2005). Modelling the effects of air pollution on health using

Bayesian Dynamic Generalised Linear Models. *University of Bath Technical Report Series* .

Lin, X. and Zhang, D. (1999). Inference in Generalised Additive Mixed Models by using Smoothing Splines. *Journal of the Royal Statistical Society Series B* **61**, 381–400.

Mar, F., Norris, G., Koenig, J. and Larson, T. (2000). Associations between Air Pollution and Mortality in Phoenix, 1995-1997. *Environmental Health Perspectives* **108**, 347–353.

Marx, B. and Eilers, P. (1998). Direct generalized additive modelling with penalized likelihood. *Computational Statistics and Data Analysis* **28**, 193–209.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models.* Chapman and Hall, 2nd edition.

Moolgavkar, S. (2000). Air Pollution and Daily Mortality in Three U.S. Counties. *Environmental Health Perspectives* **108**, 777–784.

Moolgavkar, S., Luebeck, H., Hall, T. and Anderson, E. (1995). Air Pollution and Daily Mortality in Philadelphia. *Epidemiology* **6**, 476–484.

Peng, R., Dominici, F., Pastor-Barriuso, R., Zeger, S. and Samet, J. (2005). Seasonal Analyses of Air Pollution and Mortality in 100 U.S. Cities. *American Journal of Epidemiology* **161**, 585–594.

Peng, R. D. and Welty, L. J. (2004). The NMMAPSdata package. *R News* **4**, 10–14.

Samet, J., Zeger, S., Dominici, F., Curriero, F., Coursac, I., Dockery, D., Schwartz, J. and Zanobetti, A. (2000). The National Morbidity, Mortality, and Air Pollution Study Part II: Morbidity and Mortality, from Air Pollution in the United States. *HEI Project* **96-97**, 5–47.

Schwartz, J. (1994). Nonparametric smoothing in the analysis of air pollution and respiratory illness. *The Canadian Journal of Statistics* **22**, 471–487.

Spiegelalter, D., Best, N., Carlin, B. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society series B* **64**, 583–639.

Figure 1. Time-varying effects of PM_{10} on the relative risk scale (per increase of $10\mu/m^3$), using the exact set of covariates. The actual effect (solid line) is given together with estimates from the Bayesian (dotted line) and likelihood methods (dashed line): (i) constant, (ii) seasonal, (iii), quadratic, (iv) spline.

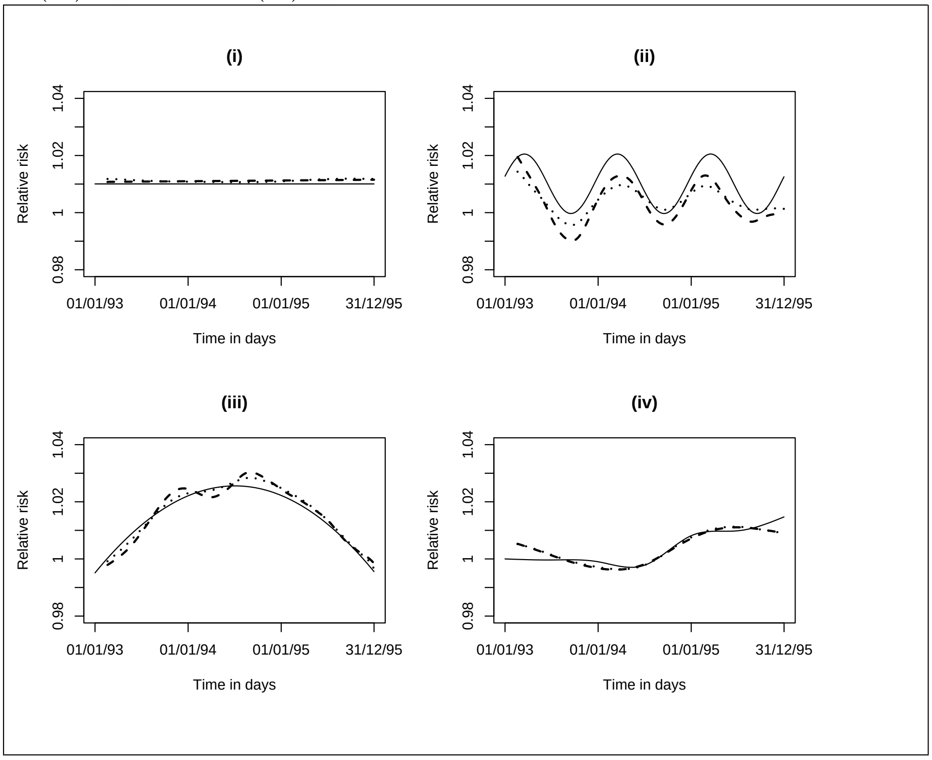

Figure 2. Time-varying effects of PM_{10} on the relative risk scale (per increase of $10\mu/m^3$), using covariates chosen by model building criteria. The actual effect (solid line) is given together with estimates from the Bayesian (dotted line) and likelihood methods (dashed line): (i) constant, (ii) seasonal, (iii), quadratic, (iv) spline.

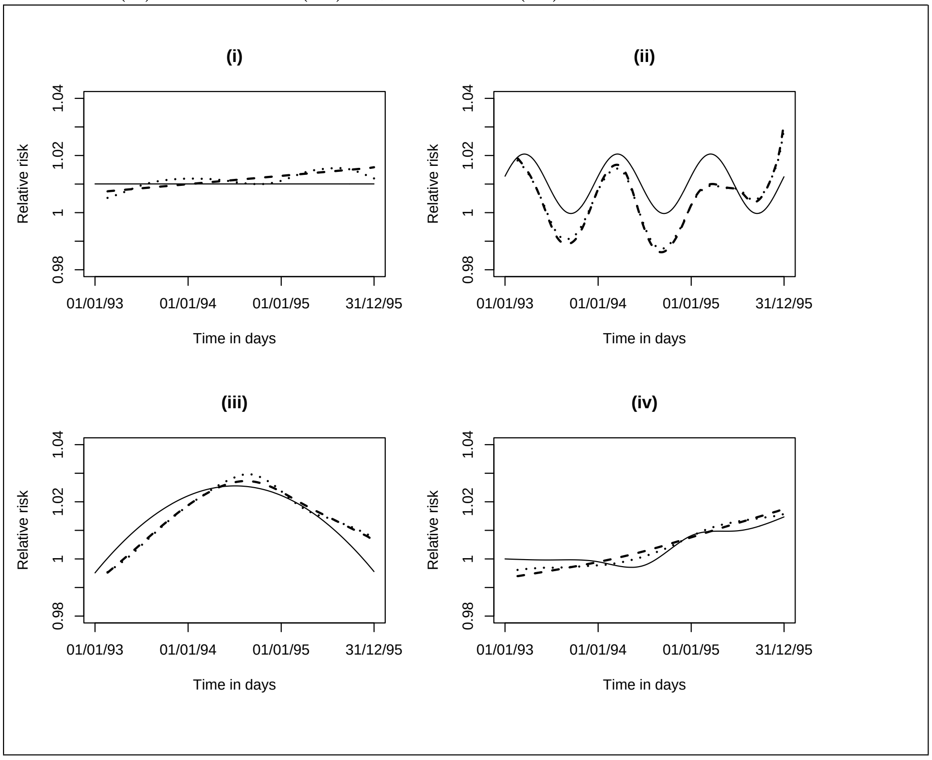

Figure 3. Counts of respiratory mortality (*) and fitted values from the Bayesian model (-).

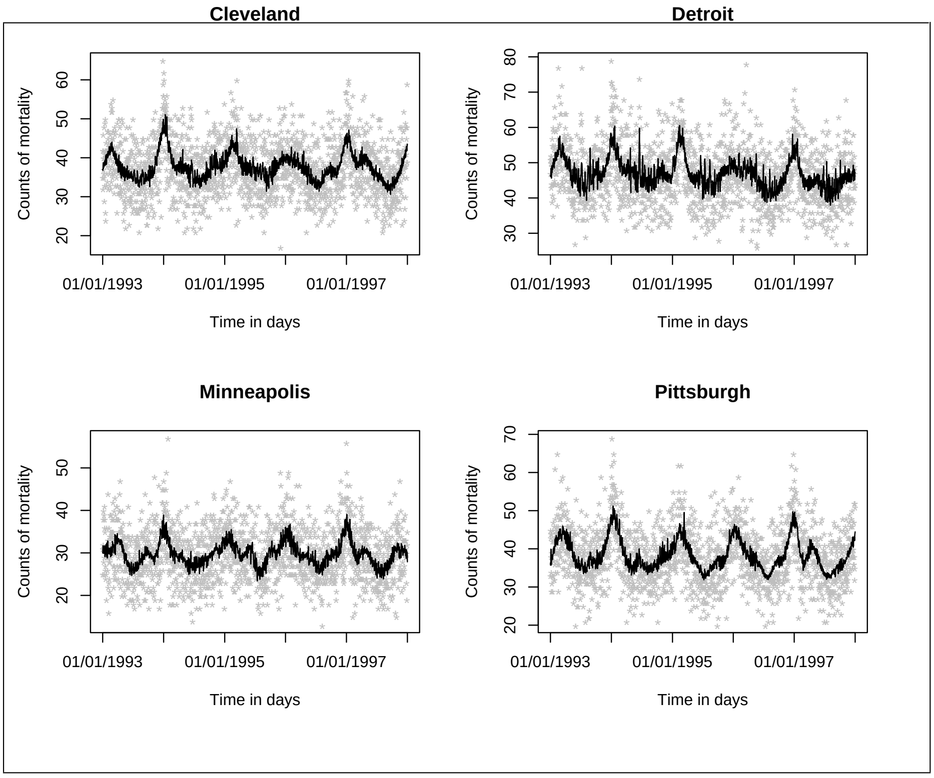

Figure 4. Time-varying coefficients for Cleveland (left) and Detroit (right). Panel (i) shows the Bayesian spline model, while panel (ii) shows the likelihood spline model. The shading represent 95% confidence / credible intervals, and the dashed line is a constant effect (model 3). Panel (iii) show the constant (dashed), seasonal (dotted) and trend models.

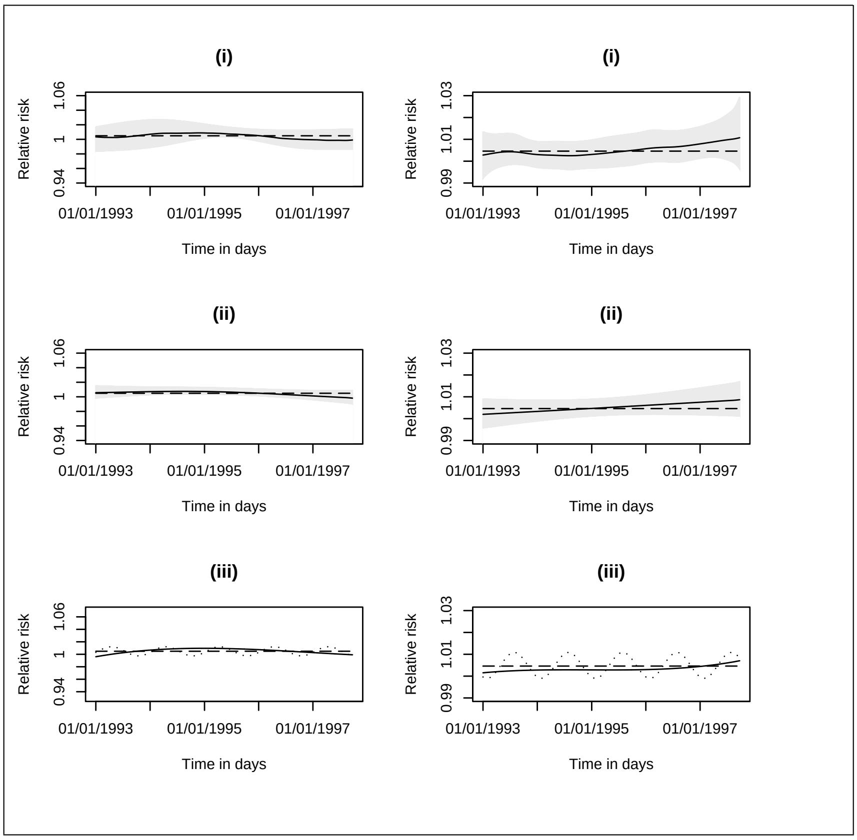

Figure 5. Time-varying coefficients for Minneapolis (left) and Pittsburgh (right). Panel (i) shows the Bayesian spline model, while panel (ii) shows the likelihood spline model. The shading represent 95% confidence / credible intervals, and the dashed line is a constant effect (model 3). Panel (iii) show the constant (dashed), seasonal (dotted) and trend models.

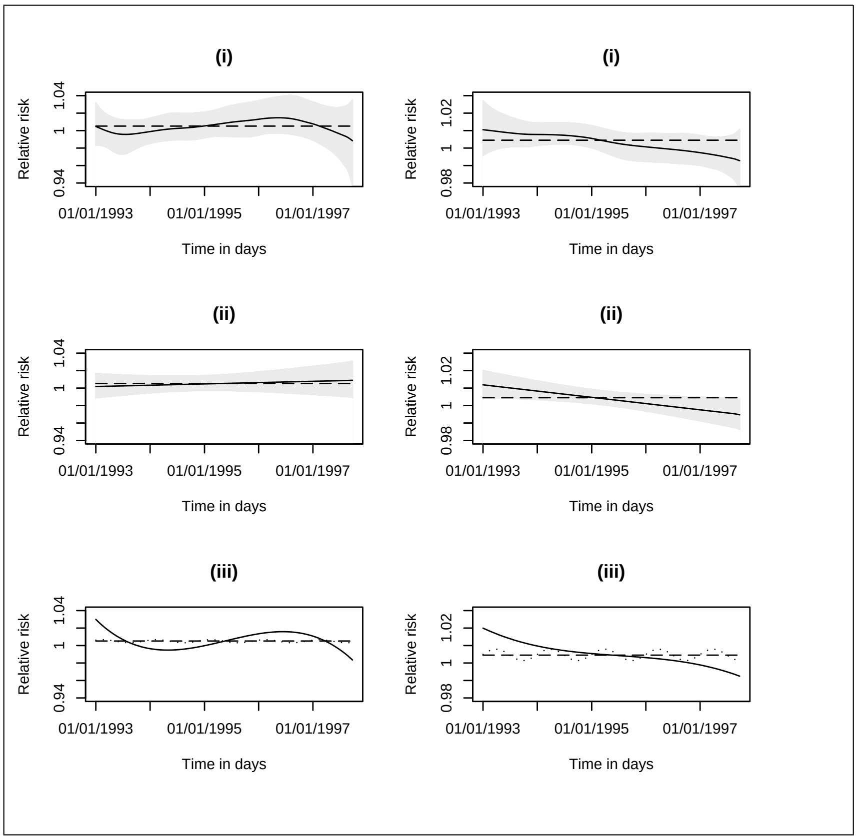

Table 1

Quantiles of the posterior distributions for τ^2 (Model 1)

City	**Posterior quantiles**				
	2.5%	25%	median	75%	97.5%
Cleveland	2.06^{-08}	3.26^{-07}	1.13^{-06}	3.92^{-06}	0.040
Detroit	1.17^{-08}	1.56^{-07}	5.32^{-07}	1.63^{-06}	1.35^{-05}
Minneapolis	2.28^{-08}	4.88^{-07}	1.69^{-06}	5.06^{-06}	2.86^{-05}
Pittsburgh	7.98^{-09}	1.78^{-07}	6.08^{-07}	1.83^{-06}	1.12^{-05}