Modelling the effects of air pollution on health using Bayesian Dynamic Generalised Linear Models

Duncan Lee and Gavin Shaddick

October 7, 2005

Abstract

The subject of the potential effects of air pollution on health is one that has received a great deal of attention in recent years, with the majority of research considering the short term, or acute, effects of changes in the level of a pollutant on subsequent counts of mortality or morbidity. In this context, there is a need for a flexible class of models that can address many of the issues encountered in modelling such associations. These include underlying seasonal patterns, serial autocorrelation and the possibility of risk changing over time. In this paper, a dynamic generalised linear model is proposed for such analyses. Such models can incorporate many of the problems encountered when using data commonly available for this type of study, i.e. daily counts of health outcome in an urban area and daily measures of pollution from ambient monitoring stations. They also provide a natural framework for incorporating many of the sources of uncertainty that may be ignored using more traditional modelling techniques. Throughout the paper, the approach is Bayesian, with particular attention to its implementation via Markov Chain Monte Carlo simulation. The potential of using such models is illustrated using daily data on respiratory deaths and particulate matter from London.

1 Introduction

The potential detrimental effects of ambient air pollution is a major issue in public health, and the subject has received a great deal of attention in recent years. A number of epidemiological studies have found consistent associations between daily levels of particulate matter (measured as PM_{10} or other similar metrics such as Black smoke (BS) or Total Suspended Particles (TSP)) and adverse health effects (mortality and morbidity), with many reporting associations at concentrations below existing guidelines and standards (see, for example, Dominici et al. (2002), Vedal et al. (2003) or Roberts (2004)). Large multi-city studies such as 'Air pollution and health: a European approach' (APHEA) (Zmirou et al. (1998)) and 'The National Morbidity, Mortality, and Air Pollution Study' (NMMAPS) (Samet et al. (2000)) have shown consistent associations between daily mortality and particulate matter across a number of US and European cities. Short-term effects of air pollution on health are estimated using time series regression models that control for seasonality and serial autocorrelation. Recently, studies have used generalised linear models (GLMs) (Daniels et al. (2004)) and generalised additive models (GAMs) (Dominici et al. (2000)) to estimate such a relationship.

This paper proposes the use of a dynamic generalised linear model (DGLM) (West et al. (1985), Fahrmeir and Tutz (2001)) to measure the effects of PM_{10} on daily mortality, using data from London. At the present time, the only known published analysis of the acute effects of air pollution on mortality using a DGLM is that of Chiogna and Gaetan (2002), who investigated the relationship between daily non-accidental mortality and PM_{10}, for residents over 65 in Birmingham Alabama. An extension to their model is proposed here by allowing the dynamic parameters to follow an autoregressive process of order p (AR[p]) rather than an AR[1] process. Analysis is implemented within a Bayesian framework using Markov Chain Monte Carlo (MCMC) simulation which is in contrast to the work of Chiogna and Gaetan (2002), who present a likelihood based analysis using the iteratively re-weighted Kalman filter and smoother (Fahrmeir and Wagenpfeil (1997)) and a cross validation criterion (Wagenpfeil (1995)). In addition a vector of static regression variables are explicitly added to the linear predictor, rather than estimating the static and dynamic variables simultaneously.

The remainder of this paper is organised as follows. Section 2 describes some of the statistical issues encountered when modelling the relationship between air pollution and mortality and compares the DGLM with regression models that are currently used. Section 3 introduces the proposed extension to the DGLM and gives details of its implementation.

An application of the proposed model is presented in Section 4, using daily mortality and PM_{10} data from London. Finally Section 5 gives a concluding discussion.

2 Modelling the effects of air pollution on mortality

2.1 Issues with modelling air pollution and mortality

There are a number of statistical issues that need to be addressed when modelling the effects of air pollution on daily mortality. A typical study will use daily data (mortality, pollution, temperature, etc.) relating to a geographical area (e.g. a city) over a specified time period. The response data are daily counts of mortality, which suggests that a Poisson model may be appropriate. However, the data can be highly overdispersed (variance greater than the mean) which invalidates the Poisson assumption of equality of mean and variance. This overdispersion is caused by the presence of unmeasured risk factors not included as covariates in the regression model. Two approaches to dealing with overdispersion are to use quasi-likelihood methods or the addition of a latent process to the model. Quasi-likelihood removes the strict distributional assumption and assumes that only the relationship between the mean and variance of the data is known, and is typically of the form, $Var(Y_t) = \phi\mathbb{E}(Y_t)$, where ϕ is a measurement of the overdispersion and can estimated from the data. The quasi-likelihood method has been extensively used in a number of different studies involving count data (Zeger (1988), Schwartz (1991) and Brumback et al. (2000)). The second method uses a latent process to model the underlying overdispersion, using a set of random effects (or a time series process that also induces serial correlation) to inflate the marginal variance whilst keeping the marginal mean unchanged (Davis et al. (2000)). If these random effects follow a Gamma distribution, this will be equivalent to modelling the data using a Negative Binomial distribution (Booth et al. (2003)), if they have a log-Gaussian distribution, it will be a simple case of a generalised linear mixed model (Breslow and Clayton (1993)).

As well as introducing overdispersion into the model, unmeasured risk factors can also introduce serial correlation between the daily counts of mortality. Standard Poisson regression models assume independence of the response data and any serial correlation present has to be explicitly modelled. Serial correlation can be modelled using parameter driven (Zeger (1988)) or observation driven (Zeger and Qaqish (1988)) models. In the former, a latent process is incorporated into the model, for example AR(p) processes or random walks of first and second order, to allow for the serial correlation. In the latter, serial correlation is induced by including functions of past response variables in the model as extra explanatory variables.

The response data (daily deaths) are highly seasonal, showing more deaths in the winter months than during the summer and may display an underlying trend. Many of the earlier studies of the acute effects of air pollution used generalised linear models with fixed parametric functions or indicator variables for each month to model trend and seasonality (Schwartz (1993) and Verhoeff et al. (1996)). The parametric functions for seasonality took the form of a pair of sine and cosine waves. Several pairs of sinusoidal terms can be included to model seasonality at different frequencies. A measure of daily temperature is often used to model seasonality though it cannot be used in isolation because it has a fixed period of one year and so cannot model seasonality with a different period. More recently, smooth functions of temperature and calender time have been used to model seasonality and trends present in the data. The smooth functions can either be non-parametric and used in a GAM setting (Dominici et al. (2002)) (such as a LOESS smoother) or fully parametric in a GLM setting (Daniels et al. (2004))(such as a natural cubic spline). Including smooth functions of 'time' and meteorological variables in the linear predictor has become the standard method for dealing with confounding in air pollution and mortality studies.

Smooth functions (natural cubic splines or LOESS) are also commonly used to model the relationship between air pollution and mortality, and there has been a lot of recent debate about the dose-response relationship (see Schwartz et al. (2001), Dominici et al. (2002) and Daniels et al. (2004)). In particular it is commonly assumed to be linear (see for example Dominici et al. (2004)), an assumption which has not been disproved by the above studies. However the linearity of the dose-response relationship is with respect to the pollution level and has no dependence on calender time. Allowing the relationship to change over time would enable any interaction with seasonality, and any long term change in pollution effect to be identified. The dynamic generalised linear model is designed to identify such changes in effects over time.

2.2 Modelling air pollution and mortality with a DGLM

A DGLM models the relationship between air pollution and mortality via a parameter that evolves dynamically over time. The evolution is governed by a stochastic process such as a random walk or a (possibly stationary) AR[p] process. A DGLM was first used in the analysis of air pollution and mortality by Chiogna and Gaetan (2002) and the general form of their model is as follows,

$$\begin{aligned}
y_t &\sim \text{Poi}(\mu_t) \qquad t = 1, \ldots, n \\
\log(\mu_t) &= \mathbf{x}_t^T \beta_t \\
\beta_t &= F\beta_{t-1} + \omega_t \qquad \omega_t \sim N_q(\mathbf{0}, Q) \\
\beta_0 &\sim N_q(a_0, Q_0)
\end{aligned} \tag{1}$$

The mortality data y_t is modelled as coming from a Poisson distribution where the mean depends on covariates $\mathbf{x}_{t_{q\times 1}}$, and parameters $\beta_{t_{q\times 1}}$ that evolve over time via an AR[1] process. The model is very general and encompasses static regression variables with constant coefficients, dynamic regression variables whose coefficients evolve over time, and latent processes such as random effects, random walks, and autoregressive processes. It can be shown that a second order random walk is equivalent to a discrete smoothing spline (Fahrmeir and Tutz (2001)), giving a connection to the commonly used GAM. Such latent processes can incorporate overdispersion and serial correlation into the model. Natural cubic splines can be included as static regression variables (just as in a GLM) to allow for non-linear confounding. The pollution effect is allowed to change over time, on a seasonal basis and in the longer term. The above definition includes as special cases, generalised linear models, parameter driven models, and observation driven models. The flexibility and broad scope of the DGLM makes it an attractive statistical model to use for the analysis of air pollution and mortality studies. This paper proposes an extension to the above definition of the DGLM and a Bayesian implementation, details of which are given in the following section.

3 Bayesian Dynamic generalised linear model

The model proposed here (equation 2) extends that of Chiogna and Gaetan (2002) because the evolution of the dynamic parameters is governed by an AR[p] rather than an AR[1] process and differs substantially in that it is implemented within a Bayesian framework. Also, a vector of static regression variables is explicitly included in the linear predictor rather than being a dynamic parameter with zero evolution variance.

The extended Bayesian dynamic generalised linear model is given below:

$$\begin{aligned}
y_t &\sim \text{Poi}(\mu_t) \qquad t = 1, \ldots, n \\
\log(\mu_t) &= \mathbf{x}_t^T \beta_t + \mathbf{z}_t^T \alpha \\
\beta_t &= F_1 \beta_{t-1} + \ldots + F_p \beta_{t-p} + \omega_t \qquad \omega_t \sim N_q(\mathbf{0}, Q) \\
\beta_0, \beta_{-1}, ..., \beta_{-p+1} &\sim N_q(a_0, Q_0) \\
\alpha &\sim N_r(m, \Sigma) \\
Q &\sim \text{Inverse-Wishart}(n_Q, S_Q^{-1})
\end{aligned} \tag{2}$$

The response data $y = (y_1, \ldots, y_n)^T_{n\times 1}$ are daily counts of mortality or morbidity, and are assumed to follow a Poisson distribution. The mean of the data is described by covariates and stochastic elements, such as random walks. Every element of response data is modelled by an $r \times 1$ vector of static covariates $\mathbf{z}_t$ and a $q \times 1$ vector of dynamic covariates or stochastic elements $\mathbf{x}_t$. The static regression variables $\mathbf{z}_t$ can be stacked in rows to give an $n \times r$ matrix of static regression variables $Z = (\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T)^T$. An $n \times q$ matrix of dynamic regression variables and known multipliers of stochastic components is given by $X = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$.

The unknown static regression parameters are denoted $\alpha = (\alpha_1, \ldots, \alpha_r)^T_{r\times 1}$. At time t the dynamic parameter $\beta_t = (\beta_{t1}, \ldots, \beta_{tq})^T_{q\times 1}$ has q elements that correspond to the q dynamic regression variables or stochastic elements. The dynamic parameters have a prior which follows an AR[p] time series process. This process is initialised by starting parameters $(\beta_{-p+1}, \ldots, \beta_0)^T_{pq\times 1}$ at times $(-p+1, \ldots, 0)$. The entire set of dynamic parameters are denoted by $\beta = (\beta_{-p+1}, \ldots, \beta_0, \beta_1, \ldots, \beta_n)^T_{(n+p)q\times 1}$. The variance of the AR[p] process is controlled by the $q \times q$ variance matrix Q. The unknown parameters are the regression parameters (β, α), and the hyperparameters from the AR[p] process $(Q, F = \{F_1, \ldots, F_p\})$. The $q \times q$ matrices $(F_1, \ldots, F_p)$ or components thereof, may be known, for example if a random walk process is assumed. The defining quantities of the priors $(m, \Sigma, n_Q, S_Q, a_0, Q_0)$ are known. The vector of daily means $\mu = (\mu_1, \ldots, \mu_n)^T_{n\times 1}$ is given by $\mu_t = \exp(\mathbf{x}_t^T \beta_t + \mathbf{z}_t^T \alpha)$.

3.1 Method of analysis

The above model (2) is implemented within a Bayesian framework using Markov Chain Monte Carlo simulation. The addition of static regression variables and the extension to an AR(p) model, requires little extra sampling effort compared with the standard DGLM (equation 1). Inference using DGLMs has been an active research topic for many years. In addition to the Bayesian simulation based approach followed here, a number of alternative approaches have been suggested. West et al. (1985) proposed an approximate analysis based on relaxing the normality of the AR[1] process, and assuming conjugacy between the data model and the AR[1] parameter model. A Bayesian setup is assumed, with linear Bayes estimation used to estimate the conditional distributions of the parameters β_t. Estimation of the unknown variance matrix Q is circumvented via the use of the discount method (Ameen and Harrison (1985)).

Fahrmeir (1992) and Fahrmeir and Wagenpfeil (1997) proposed a classical analysis by estimating the dynamic parameters by the mode of the joint posterior $p(\beta_1, \ldots, \beta_n | y)$, using the Kalman filter and a Fisher scoring step, details of which can be found in Meinhold and Singpurwalla (1983) and McCullagh and Nelder (1989) respectively. An iterative algorithm simultaneously estimates the dynamic parameters β_t and the variance matrix Q, by alternately re-estimating each based on the current estimate of the other. The dynamic parameters are estimated using the iteratively re-weighted Kalman filter and smoother, while the matrix Q can be estimated using the expectation-maximisation (EM) algorithm. Other approaches include approximating the posterior density by piecewise linear functions (Kitagawa (1987)), and using numerical integration to estimate the parameters (Fruhwirth-Schnatter (1994)).

For the AR[p] DGLM (equation 2), the joint posterior distribution of all the parameters is given by:

$$
\begin{aligned}
p(\beta, \alpha, Q, F|y) &= p(y|\beta, \alpha, Q, F)p(\beta, \alpha, Q, F) \\
&= p(y|\beta, \alpha)p(\alpha)p(\beta|Q, F)p(Q)p(F) \\
&= \prod_{t=1}^{n} p(y_t|\beta_t, \alpha)p(\alpha) \prod_{t=1}^{n} p(\beta_t|\beta_{t-1}, \ldots, \beta_{t-p}, Q, F)p(\beta_{-p+1}) \ldots p(\beta_0)p(Q)p(F)
\end{aligned} \tag{3}
$$

The parameters of the AR(p) process (Q, F) have full conditionals that are standard distributions. As a result the full conditionals can be directly sampled from using a Gibbs step. The blocks of regression parameters (β, α) have non-standard full conditionals, so the full conditionals cannot be sampled from directly. A number of different sampling schemes have been suggested in the literature, especially for the dynamic parameters β (see for example Shephard and Pitt (1997), Gamerman (1998) and Knorr-Held (1999)). The following section (3.2) gives an overview of the sampling algorithm, with specific details of a selection of the sampling schemes given in sections (3.2.1-3.2.3).

3.2 Overall sampling algorithm

The proposed sampling scheme is a hybrid Metropolis within Gibbs algorithm (Gelman et al. (2003)) and after calculating starting values for $(\beta^0, \alpha^0, Q^0, F^0)$, the parameters are repeatedly sampled in the following blocks;

(a) **Dynamic parameters** $\beta = (\beta_{-p+1}, \ldots, \beta_1, \ldots, \beta_n)$.
The full conditional of $\beta = (\beta_{-p+1}, \ldots, \beta_n)$ is non-standard because it is a product of n Gaussian AR[p] prior distributions and Poisson data. Further details are given in Section (3.2.1).

(b) **Static parameters** $\alpha = (\alpha_1, \ldots, \alpha_r)$.
The full conditional is again non-standard, being the product of Gaussian and Poisson distributions, further details of the sampling scheme are given in Section (3.2.2).

(c) **Variance matrix** Q.
The prior for Q is chosen to be an Inverse-Wishart(n_Q, S_Q^{-1}) distribution because it

is conjugate to the Gaussian AR[p] model. The full conditional for Q is an Inverse-Wishart distribution and can be directly sampled using a Gibbs sampling step.

(d) **AR(p) matrices**, $F = (F_1, \ldots, F_p)$ (or components of).
The full conditional for F will depend on the form of the AR(p) process and the dimension of β_t (number of dynamic variables). The most common types will be the univariate AR[1] and AR[2] processes for which details are given in Section (3.2.3). In these cases F_i is sampled from its Gaussian full conditional distribution using a Gibbs sampling step.

3.2.1 Simulation from $\mathbf{p}(\beta|y, \alpha, Q, F)$

The full conditional of $\beta = (\beta_{-p+1}, \ldots, \beta_n)$ is a product of n Poisson observations and a Gaussian AR[p] prior. The prior is a product of n AR[p] multivariate Gaussian distributions for β_t $t = 1, \ldots, n$ and p multivariate Gaussian distributions for the initialising parameters $\beta_{-p+1}, \ldots, \beta_0$. The full conditional for β is given by:

$$\begin{aligned} p(\beta|y, \alpha, Q, F) &\propto p(\beta) \prod_{t=1}^{n} p(y_t|\alpha, \beta_t) \\ &\propto N(\beta_{-p+1}|a_0, Q_0) \ldots N(\beta_0|a_0, Q_0) \prod_{t=1}^{n} N(\beta_t|F_1\beta_{t-1} + \ldots + F_p\beta_{t-p}, Q) \\ &\times \prod_{t=1}^{n} \mathrm{Poi}(y_t|\mu_t(\alpha, \beta_t)) \end{aligned} \quad (4)$$

The full conditional is non-standard and cannot efficiently be sampled from directly. Rejection sampling combined within the Gibbs step has been suggested (Fahrmeir et al. (1992)), but very low acceptance rates were found, which made the algorithm very slow. A number of Metropolis-Hastings algorithms have been suggested, which sidestep the rejection sampling problem including using proposal distributions based on Fisher scoring steps and Taylor expansions (Shephard and Pitt (1997), Gamerman (1998)). These methods can be computationally expensive, and the proposal used here is that suggested by Knorr-Held, which bases the proposal density on the prior of the data (Knorr-Held (1999)). The proposals are computationally cheap to generate compared with the Fisher scoring and Taylor expansions required by the methods mentioned above, and the Metropolis-Hastings acceptance rate is a simple form and easy to calculate. The parameter vector β is sampled in blocks Chib and Greenberg (1995) which is a sensible intermediate strategy between the low acceptance rates obtained from sampling high dimensional vectors, and the computational burden of sampling variables individually. The block size g is a tuning parameter to achieve the desired acceptance rates. Knorr-Held suggests that the algorithm is optimal when the acceptance rates are well below 1. The proposal $q()$ for all the blocks is based on the prior for β which can be re-written as a multivariate Gaussian distribution with a singular variance matrix:

$$\begin{aligned} p(\beta|F,Q) &= N(\beta_{-p+1}|a_0,Q_0)\dots N(\beta_0|a_0,Q_0)\prod_{t=1}^{n} N(\beta_t|F_1\beta_{t-1}+\dots+F_p\beta_{t-p},Q) \\ &\propto \exp\left(-\frac{1}{2}\beta^T K\beta\right) \end{aligned} \tag{5}$$

The prior is not a proper multivariate Gaussian distribution because the precision matrix K is singular, so the variance matrix of the prior β does not exist. The matrix K is square and has dimensions $(n+p)q \times (n+p)q$ and is defined in terms of $q \times q$ blocks K_{tt}:

$$K = \begin{bmatrix} K_{-p+1,-p+1} & \dots & K_{-p+1,n} \\ \vdots & & \vdots \\ K_{n,-p+1} & \dots & K_{n,n} \end{bmatrix} \tag{6}$$

The exact form of the matrix K depends on the order p of the AR[p] process, and the matrix has a bandwidth of p blocks (blocks further than p blocks from the diagonal are all zero). As an example, the matrix K is given below for a first order autoregressive process. Formulae for higher order autoregressive processes are similar. All blocks in K other than the ones stated below are zero.

First order

$$
\begin{aligned}
K_{tt} &= \begin{cases} F_1^T Q^{-1} F_1 + Q_0^{-1} & t = 0 \\ F_1^T Q^{-1} F_1 + Q^{-1} & t = 1, \ldots, n-1 \\ Q^{-1} & t = n \end{cases} \\
K_{t,t+1} &= -F_1^T Q^{-1} \qquad \forall\ t \\
K_{t,t-1} &= -Q^{-1} F_1 \qquad \forall\ t
\end{aligned}
$$

The proposal density $q()$ used to update a block of parameters β_{rs}, is the multivariate Gaussian conditional distribution for that block β_{rs} given all the other elements of the parameter vector β denoted β_{-rs}. For iteration j the proposal β^*_{rs} is drawn from the density given below:

$$
q[\beta_{rs}^{j-1}, \beta_{rs}^*] = p(\beta_{rs} | \beta_{-rs}, F, Q) \ \sim \ N(\mu_{rs}, \Sigma_{rs}) \tag{7}
$$

The mean and variance can be obtained by using standard properties of the multivariate Gaussian distribution.

$$
\begin{aligned}
\mu_{rs} &= \left[\begin{array}{ll} -K_{rs}^{-1} K_{-p+1,r-1} \beta_{-p+1,r-1} & \text{if s=n} \\ -K_{rs}^{-1} K_{s+1,n} \beta_{s+1,n} & \text{if r=-p+1} \\ -K_{rs}^{-1} (K_{-p+1,r-1} \beta_{-p+1,r-1} + K_{s+1,n} \beta_{s+1,n}) & \text{otherwise} \end{array} \right. \\
\Sigma_{rs} &= K_{rs}^{-1}
\end{aligned}
$$

To avoid any mixing problems at the boundaries of each block, the first block will be of random length as suggested by Knorr-Held (Knorr-Held (1999)). The acceptance probability of a move from β_{rs}^{j-1} to β^*_{rs} is simple to calculate and is given by:

$$
\min \left\{ 1, \frac{\prod_{t=r}^{s} \text{Poi}(y_t | \mu_t(\beta_t^*, \alpha^{j-1}))}{\prod_{t=r}^{s} \text{Poi}(y_t | \mu_t(\beta_t^{j-1}, \alpha^{j-1}))} \right\} \tag{8}
$$

Details can be found in Knorr-Held (Knorr-Held (1999)).

3.2.2 Simulation from $\mathbf{p}(\alpha|y, \beta, Q, F)$

The full conditional of the vector of static regression parameters, α, is the product of a Gaussian prior and Poisson data and is thus non standard.

$$p(\alpha|y, \beta, Q) \;\propto\; N(\alpha|m, \Sigma) \prod_{t=1}^{n} \mathrm{Poi}(y_t|\mu_t(\alpha, \beta_t)) \tag{9}$$

A Gaussian prior is chosen because prior information is simple to specify in that form, but it does lead to a non-standard posterior distribution and cannot be sampled from directly, however it can be sampled from using a Metropolis-Hastings step. Two popular choices for the proposal distribution are random walk and Fisher scoring proposals (Fahrmeir and Tutz (2001)). The random walk proposal is chosen because of its computational cheapness compared with calculating a Fisher scoring step at each iteration and because the Fisher scoring proposal does not have a tuning parameter, whereas the variance of the random walk can be tuned to give good acceptance rates. The vector α could be updated in its entirety during a single Metropolis step, updated in blocks using a block Metropolis algorithm (Chib and Greenberg (1995)), or each element could be updated singularly using a univariate random walk proposal. The block Metropolis algorithm was chosen because it was a compromise between the high acceptance rates of univariate sampling, and the speed of variable generation (Liu et al. (1994)) and improved mixing that arises when sampling large blocks of parameters together. The proposal density for a block of parameters α_{rs} for iteration j is given by:

$$q[\alpha_{rs}^{j-1}, \alpha_{rs}^{*}] \;\sim\; N(\alpha_{rs}^{j-1}, \Upsilon) \tag{10}$$

The variance matrix Υ is diagonal with elements that can be tuned to give good acceptance rates.

3.2.3 Simulation from $\mathbf{p}(F|y, \beta, \alpha, Q)$

The full conditional for F will depend on the form of the AR[p] process and the dimension of β_t (number of dynamic variables). The most common types will be the univariate AR[1]

($\beta_t = \phi\beta_{t-1} + \omega_t$) and AR[2] ($\beta_t = \phi_1\beta_{t-1} + \phi_2\beta_{t-2} + \omega_t$) processes. In the case of the AR(1) process, the unknown parameter, (ϕ), is assigned a flat prior and will replace the matrices F (in equation (2)). The full conditional for ϕ in the AR[1] process takes the form of a Gaussian distribution with mean, $\mu = \frac{\sum_{t=1}^{n}\beta_t\beta_{t-1}}{\sum_{t=1}^{n}\beta_{t-1}^2}$, and variance, $\sigma^2 = \frac{Q}{\sum_{t=1}^{n}\beta_{t-1}^2}$. In this case Q becomes a scalar variance component. Similar results can be found for AR[2] process.

4 Application

The model described in Section 3 is illustrated by analysing the relationship between air pollution and mortality in Greater London.

4.1 Data

The data relate to daily observations from the Greater London area during the period 1$^{\text{st}}$ January 1995 until 31$^{\text{st}}$ December 1997, although due to the choice of lags for explanatory variables (See sections 4.2.1 - 4.2.2) the data starts from the 4$^{\text{th}}$ January 1995. The response data comprise daily counts of respiratory mortality for all ages and are shown in Figure 1. The data show a strong seasonal pattern with a large increase in the number of deaths during the winter of 1996/1997. The cause of this peak is unknown and research has shown no influenza epidemic during this time (which has been associated with large increases of this type (Griffin and Neuzil (2002)). Previous analysis of data from this period in London has shown little spatial variation in PM_{10} (Shaddick and Wakefield (2002)), so the PM_{10} data are daily averages taken across eleven automatic monitoring sites in London. The daily mean temperature (^{0}C) during the study period, measured at Heathrow airport, was included as previous studies have shown climate variables to be important confounders in air pollution and mortality studies (e.g. Dominici et al. (2000), Gwynn et al. (2000), Schwartz et al. (2001)).

4.2 Statistical Model

The final statistical model is given below, justifications for this choice of model are given in the following subsections (4.2.1-4.2.3).

$$
\begin{aligned}
y_t &\sim \text{Poi}(\mu_t) \\
\log(\mu_t) &= \gamma_t + \beta_t PM_{10}^{av4} + \alpha_1 + S(temp^{av4}, 6) + S(time, 21) \\
\gamma_t &= \gamma_{t-1} + \omega_t \qquad \omega_t \sim N(0, W) \\
\gamma_0 &\sim N(0, 1000) \\
\beta_t &= \phi\beta_{t-1} + \nu_t \qquad \nu_t \sim N(0, Q) \\
\beta_0 &\sim N(0, 1000) \\
\alpha &\sim N(m, \Sigma) \\
Q &\sim \text{Inverse Gamma}(0.1, 0.1) \\
W &\sim \text{Inverse Gamma}(0.1, 0.1)
\end{aligned} \tag{11}
$$

Where $S(v, df)$ denotes a natural cubic spline applied to variable v with df degrees of freedom. The variable 'time' refers to calender time and 'temp' refers to daily mean temperature. The parameters of the two natural spline terms are estimated with the fixed intercept α_1 as static parameters, and are collectively denoted α.

4.2.1 Controlling for seasonality and auto-correlation

The model building strategy was to first fit a model with confounding variables only, to model any long-term trends and seasonality, before assessing the effect of PM_{10}. The confounding variables were initially fitted as static variables (denoted α in the model definition 2) using a generalised linear model (GLM). There was little evidence of a long-term linear trend in the data with a non-significant parameter estimate very close to zero (0.00008), and as a result the linear trend term was removed from the model. To control for seasonality, smooth functions of calender time and daily mean temperature were fitted to the data. Smooth functions were chosen because they allow the effects of seasonality to vary from year to year, where as fixed parametric alternatives such as pairs of sine and cosine terms, do not allow such variability. Natural cubic splines were chosen because of their parametric setting and their use in other recent studies (Daniels et al. (2004)). Following the work of Dominici et al (Dominici et al. (2000)) the spline of calender time was fitted with 7 degrees of freedom for each year giving 21 degrees of freedom over the three year period of the data. Six degrees of freedom were chosen for the spline of temperature because it allows for 2 degrees of freedom for each year of the data. Different lags and moving averages of temperature were tried in the model, the most significant of which turned out to be a 4-day average of the current day and the previous 3 days (denoted temp^{av4}).

After allowing for seasonality, a constrained first order random walk was added to the linear predictor to allow for any serial autocorrelation and overdispersion present in the data. The random walk term (denoted γ_t) acted as a moving intercept, the sum of which was constrained to sum to zero. The constraint was enforced to help identifiability of the parameters once a dynamic parameter for PM_{10} was introduced into the model. At first, AR[1] and AR[2] processes were fitted but there were problems with convergence of the MCMC sampling algorithm. In the AR[1] model, the majority of the simulations of the AR[1] parameter were greater than one, so a non-stationary random walk process was used.

4.2.2 The effects of PM_{10}

Finally, the effect of PM_{10} was assessed, and a number of different lags and moving averages were fitted. A 4 day average of the current and previous 3 days pollution proved to be the most significant and thus is included in the model (Lumley and Sheppard (2000)). Both AR[1] and AR[2] processes were fitted to the data, but in the simulation of the AR[2] model the AR[2] parameter converged to zero, and so an AR[1] process was fitted to describe the evolution of the PM_{10} coefficient over time. The two dynamic elements (constrained random walk intercept term and the PM_{10} coefficient) are chosen to be independent because they are designed to model different aspects of the data. The constrained random walk is designed to allow for serial correlation and over-dispersion and not to influence the magnitude of the fitted values, whereas the AR[1] process for the PM_{10} coefficient models the (possibly non-linear relationship) between air pollution and mortality. The special case of a constant effect of PM_{10} can be estimated from this model if $\phi \approx 1$ and $Q \approx 0$.

4.2.3 Choice of priors

Non-informative priors were chosen for the independent evolution variances Q and W using the Inverse Gamma family (identical to the univariate Inverse Wishart distribution) because of its conjugacy with Gaussian variance parameters. A number of different Inverse-Gamma$(\varepsilon, \varepsilon)$ priors were tested but changing ε had little effect on the results. A flat prior was chosen for the AR[1] parameter ϕ because the data were highly informative about its value. Non-informative Gaussian priors were chosen for β_0 and γ_0 because they are conjugate to the AR[1] and first order random walk Gaussian priors assumed for the rest of the dynamic elements. The prior for the static parameters (α) was a multivariate Gaussian distribution, with a mean based on estimates from a generalised linear model of similar data from Greater London from years previous to the period of study. The variance of this prior was chosen to be diagonal with large diagonal elements (1000) to ensure it was non-informative.

4.3 Results

The above model (11) was run using 2 parallel chains, where starting points were random samples from overdispersed versions of the priors (e.g. Student-t distributions replacing Gaussian distributions). There are a large number of parameters (1094 elements $\beta = (\beta_0, \dots, \beta_{1093})$, another 1094 elements $\gamma = (\gamma_0, \dots, \gamma_{1093})$, Q, W, ϕ and 28 static parameters $\alpha = (\alpha_1, \dots, \alpha_{28})$) in the model and, to aid convergence, explanatory variables (PM_{10} and the basis functions for the natural cubic splines) were standardised to have a mean of zero and a standard deviation of one before inclusion in the model (and were subsequently back-transformed when obtaining results from the posterior distribution. Convergence was monitored using the methods of Gelman et al. (2003). A burn in of 50,000 iterations was run, by which point convergence was assessed to have been reached, and a further 100,000 iterations were then completed for each chain, which were then thinned by 5 to reduce autocorrelation. This resulted in the two chains each contributing 20,000 samples for the analysis. The fitted values μ_t^s were calculated for each day $t = 1, \dots, 1093$ and for each sample $s = 1, \dots, 40,000$ as:

$$\mu_t^s \;=\; \exp(\gamma_t^s + \beta_t^s PM_{10}^{av4} + \alpha_1^s + S^s(temp^{av4}, 6) + S^s(time, 21)) \qquad (12)$$

Where $S^s()$ denotes the simulation of parameters that make up the spline from sample s. A plot of the respiratory mortality data and the median values from the fitted model are shown in Figure 1. The fitted model estimates the overly large spike in the number of deaths in January 1997 well, in particular correctly estimating different numbers of deaths during the winters of 1996 and 1997. However accurate estimation of the outlying data point in January 1997 (low number of deaths for a January day) appears to be spurious, and simply a result of the number of parameters in the model.

The three panels of Figure 2 give diagnostic plots for the results of the model. Figure 2(a) shows a plot of the Pearson residuals which show no pattern, suggesting that confounding was adequately controlled for. The autocorrelation function (ACF) and partial autocorrelation function (PACF) for the residuals are shown in Figure 2(b) and (c) and show little evidence of residual correlation. The first order random walk (denoted γ_t) and the natural cubic splines of calendar time and temperature appear to have successfully removed the abundant serial correlation present in the respiratory mortality data (ACF and PACF not shown).

Table 1 gives summary statistics of the marginal posterior distribution of the hyperparameters. The posterior median of ϕ is 0.33 resulting in the coefficient for PM_{10} evolving as a stationary AR[1] process. This is in contrast to the work of Chiogna and Gaetan (2002) who modelled the evolution of the PM_{10} coefficient as a non-stationary first order random walk. The posterior median of the evolution variance Q, 0.042 (standard deviation of approximately 0.2) is large enough to allow significant change in the effect of PM_{10} over time. In contrast, the evolution variance of the first order random walk that controls for serial autocorrelation is small 0.00023 (standard deviation of 0.014). This suggests that the majority of the serial correlation present in the response data may have been successfully modelled by the splines that control for confounding. A plot of the simulated median of the random walk γ_t over time is given in Figure 3 and shows a seasonal pattern that is removed and so does not confound the results. By design the random walk is constrained so that the sum of it's terms is zero. This ensured that it played little role in estimating the absolute size of the fitted values.

The interesting part of the model was that it allowed the effect of PM_{10} to vary over time according to a first order autoregressive process. A plot of the smoothed median of the time varying coefficient β_t^{med} of PM_{10} with a pointwise 95% credible interval is given in Figure 4. The median β_t^{med} was highly variable over the range of plausible PM_{10} effects because the estimate (median) of the variance Q, that controls the variability of the AR[1] process for β_t, was not close to zero (standard deviation 0.2). As a result the median and the limits of the credible interval were smoothed so that the graph shows the underlying trend in the effects of PM_{10}. The estimates of the effects of PM_{10} ranged from -0.0018 to 0.006 over time and showed a larger effect of air pollution during 1997. The overall median effect of air pollution for the three year period is 0.002. The graph shows that although there does appear to be a long-term trend in the effects of PM_{10} over time, the credible intervals are too wide to remove the possibility of a constant effect over the study period. In contrast the classical pointwise confidence interval (represented by dark grey shading on figure 4) for the smoothed estimate is much narrower and does not include the variability induced by estimating the median effects series.

5 Discussion

This paper proposes a model for assessing the relationship between PM_{10} and daily respiratory mortality which allows for the possibility that the effect of PM_{10} may change over time. It also provides a procedure for incorporating the possibility of overdispersion and serial correlation, after allowing for the effects of seasonality.

The majority of PM_{10} and mortality analyses estimate a single constant effect of PM_{10} for the entire study period. The results from this model estimate a series of effects of PM_{10} one for each day of the study. The smoothness of these effects is controlled by assuming they follow an autoregressive process. In this study the effects of PM_{10} showed little change over the first two years, and then a marked increase in 1997, although the credible intervals are wide enough not to exclude the possibility of a constant effect. The width of the credible intervals reflects the fact that the model incorporates much more of the inherent variability than has been the case when using more traditional methods, and is the result of the Bayesian framework adopted. In recent studies, investigators have examined the possibility of non-linear effects of pollution with respect to the actual level of pollution. This could have been a reason behind the changing effects over time, i.e. if temporal increases in the effect were associated with increased levels of PM_{10}. However an examination of the size of effect with relation to the magnitude of the pollution level showed this not to be the case. The apparent change in the effect of PM_{10} may therefore be due to seasonal or long-term differences in its composition.

The overall relative risk for an increase of 10 μm ranged from 0.98 to 1.06 (over time) with the overall median value being 1.02. It is noticeable that while the values for the relative risks are of similar magnitude to other studies, the credible intervals (based on the values of β in figure 4) are wider, as they incorporate much of the inherent variability that would otherwise be ignored.

The model presented here is flexible and extends previous work by allowing the dynamic effects of PM_{10} to follow an AR[p] process, however the analysis of the London data showed no evidence that more than an AR[1] process was required. The AR[1] process estimated was stationary (with an AR[1] parameter of 0.33) which is noticeably different from the random walk used by Chiogna and Gaetan (2002).

Allowing the effects of air pollution to vary over time merits further investigation especially pollutants such as PM_{10} whose composition may depend upon the season. However it may be desirable to consider alternative methods for controlling the 'smoothness' of the effects over time. The first order autoregressive process allowed the effects to have a large amount of variability and a smoother change over time may be more realistic. One possibility that may merit further investigation would be to control the changing effects of air pollution using a smooth function such as a penalised spline.

Acknowledgements

The mortality data was provided by the Small Area Health Statistics Unit, which is funded by grants from the Department of Health, Department of the Environment, Food and Rural Affairs, Health and Safety Executive, Scottish Executive, National Assembly of Wales, and Northern Ireland Assembly.

References

Ameen, J. and P. Harrison (1985). Normal Discount Bayesian Models. *Bayesian Statistics 2 2*, 271–198.

Booth, J., G. Casella, H. Friedl, and J. Hobert (2003). Negative Binomial Loglinear Mixed Models. *Statistical Modelling 3*, 179–191.

Breslow, N. and D. Clayton (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association 88*, 9–25.

Brumback, B., L. Ryan, J. Schwartz, L. Neas, and P. Stark (2000). Transitional Regression Models, with Application to Environmental Time Series. *Journal of the American Statistical Association 95*, 16–27.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician 49*, 327–335.

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

Daniels, M., F. Dominici, S. Zeger, and J. Samet (2004). The National Morbidity, Mortality, and Air Pollution Study Part III: Concentration-Response Curves and Thresholds for the 20 Largest US Cities. *HEI Project 96-97*, 1–21.

Davis, R., W. Dunsmuir, and Y. Wang (2000). On autocorrelation in a Poisson regression model. *Biometrika 87*, 491–505.

Dominici, F., M. Daniels, S. Zeger, and J. Samet (2002). Air Pollution and Mortality: Estimating Regional and National Dose-Response Relationships. *Journal of the American Statistical Association 97*, 100–111.

Dominici, F., A. McDermott, and T. Hastie (2004). Improved Semiparametric Time Series Models of Air Pollution and Mortality. *Journal of the American Statistical Association 99*, 938–948.

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

Fahrmeir, L. (1992). Posterior Mode Estimation by Extended Kalman Filtering for Multivariate Dynamic Generalized Linear Models. *Journal of the American Statistical Association 87*, 501–509.

Fahrmeir, L., W. Hennevogl, and K. Klemme (1992). Smoothing in dynamic generalized linear models by Gibbs sampling. *Advances in GLIM and Statistical Modelling*, 85–90.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Modles* (2nd ed.). Springer.

Fahrmeir, L. and S. Wagenpfeil (1997). Penalized likelihood estimation and iterative Kalman smoothing for non-Gaussian dynamic regression models. *Computational Statistics and Data Analysis 24*, 295–320.

Fruhwirth-Schnatter, S. (1994). Applied state space modelling of non-Gaussian time series using integration-based Kalman filtering. *Statistics and Computing 4*, 259–269.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalized linear models. *Biometrika 85*, 215–227.

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

Griffin, M. and K. Neuzil (2002). The Global Implications of Influenza in Hong Kong. *The New England Journal Of Medicine 347*, 2159–2162.

Gwynn, R., R. Burnett, and G. Thurston (2000). A Time-Series Analysis of Acidic Particulate Matter and Daily Mortality and Morbidity in the Buffalo, New York Region. *Environmental Health Perspectives 108*, 125–133.

Kitagawa, G. (1987). Non-Gaussian State-Space Modelling of Nonstationary Time Series. *Journal of the American Statistical Association 82*, 1032–1041.

Knorr-Held, L. (1999). Conditional Prior Proposals in Dynamic Models. *Scandinavian Journal of Statistics 26*, 129–144.

Liu, J., W. Wong, and A. Kong (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika 81*, 27–40.

Lumley, T. and L. Sheppard (2000). Assessing seasonal confounding and model selection bias in air pollution epidemiology using positive and negative control analyses. *Environmetrics 11*, 705–717.

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

Meinhold, R. and N. Singpurwalla (1983). Understanding the Kalman Filter. *The American Statistician 37*, 123–127.

Roberts, S. (2004). Biologically Plausible Particulate Air Pollution Mortality Concentration-Response Functions. *Environmental Health Perspectives 112*, 309–313.

Samet, J., S. Zeger, F. Dominici, F. Curriero, I. Coursac, D. Dockery, J. Schwartz, and A. Zanobetti (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. (1991). Particulate Air Pollution and Daily Mortality in Detroit. *Environmental Research 56*, 204–213.

Schwartz, J. (1993). Air Pollution and Daily Mortality in Birmingham, Alabama. *American Journal of Epidemiology 137*, 1136–1147, abstract =.

Schwartz, J., F. Ballester, M. Saez, P. Perez-Hoyos, J. Bellido, K. Cambra, F. Arribas, A. Canada, M. Jose Perez-Boillos, and J. Sunyer (2001). The Concentration-Response Relation between Air Pollution and Daily Deaths. *Environmental Health Perpespectives 109*, 1001–1006.

Shaddick, G. and J. Wakefield (2002). Modelling multiple pollutants and multiple sites. *Appl. Statist. 51*, 351–372.

Shephard, N. and M. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika 84*, 653–667.

Vedal, S., M. Brauer, R. White, and J. Petkau (2003). Air Pollution and Daily Mortality in a City with Low Levels of Pollution. *Environmental Health Perpespectives 111*, 45–51.

Verhoeff, A., G. Hoek, J. Schwartz, and J. van Wijnen (1996). Air Pollution and Daily Mortality in Amsterdam. *Epidemiology 7*, 225–230.

Wagenpfeil, S. (1995). Hyperparameter Estimation in Exponentially Family State Space Models. *DFG Sonderforschungsbereich 386.*

West, M., J. Harrison, and H. Migon (1985). Dynamic Generalized Linear Models and Bayesian Forecasting. *Journal of the American Statistical Association 80*, 73–83.

Zeger, S. (1988). A regression model for time series of counts. *Biometrika 75*, 621–629.

Zeger, S. and B. Qaqish (1988). Markov Regression Models for Time Series: A Quasi-Likelihood Approach. *Biometrics 44*, 1019–1031.

Zmirou, D., J. Schwartz, M. Saez, A. Zanobetti, B. Wojtyniak, G. Touloumi, C. Spix, A. Ponce de Leon, Y. Le Moullec, L. Bacharova, J. Schouten, A. Ponka, and K. Katsouyanni (1998). Time-Series Analysis of Air Pollution and Cause Specific Mortality. *Epidemiology 9*, 495–503.

Figure 1: **Counts of respiratory mortality (*) and the fitted model (-)**

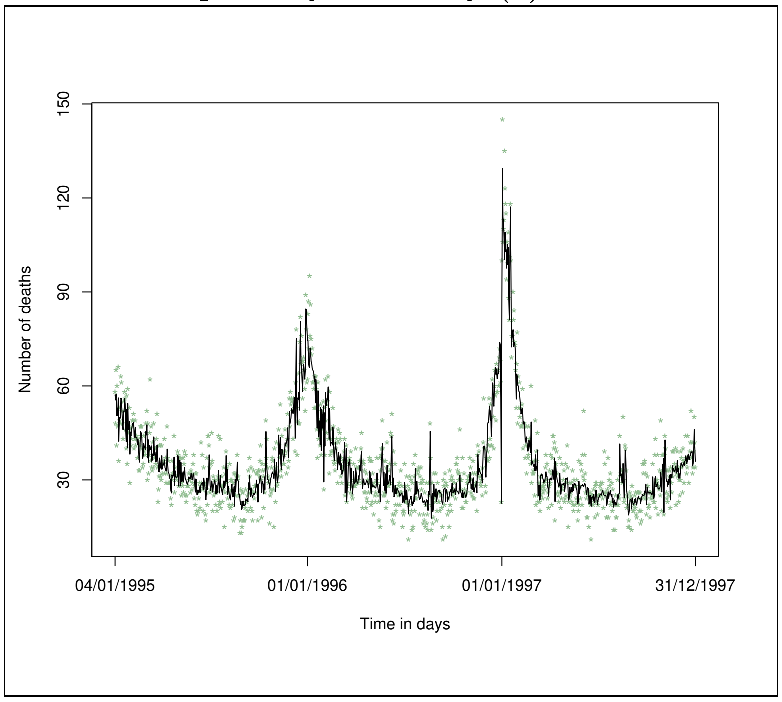

Table 1: **Summary of the posterior distributions of the variance and AR[1] parameters**

Hyperparameter	**Posterior quantiles**				
	2.5%	25%	median	75%	97.5%
Q	0.03403	0.03867	0.04128	0.04403	0.04995
W	0.0002025	0.0002168	0.0002303	0.0002446	0.0002628
ϕ	0.1712	0.2781	0.3308	0.3830	0.4772

Figure 2: **Pearson residuals from the fitted model: (a) residuals against time, (b) their ACF and (c) PACF**

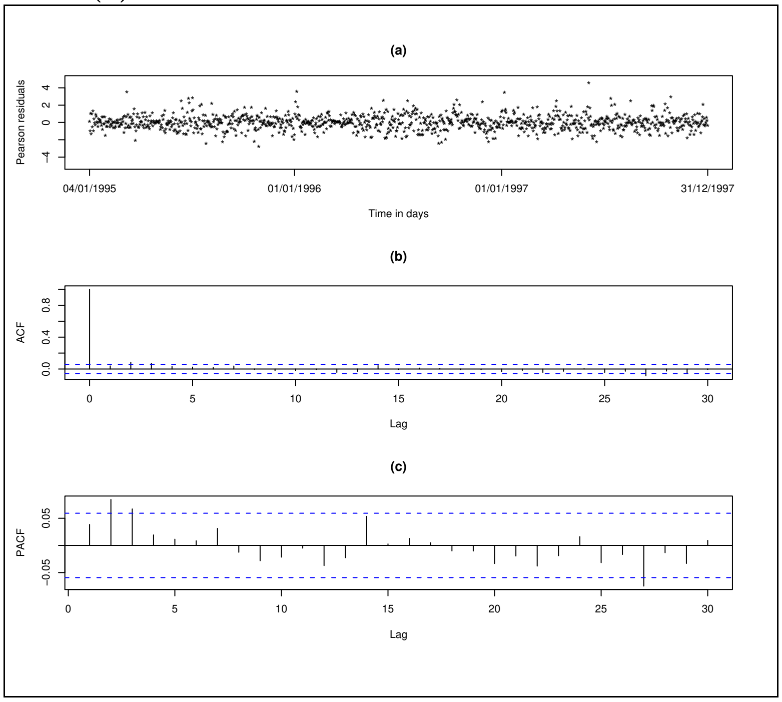

Figure 3: **The first order random walk (RW[1]) (γ_t) that allows for serial correlation**

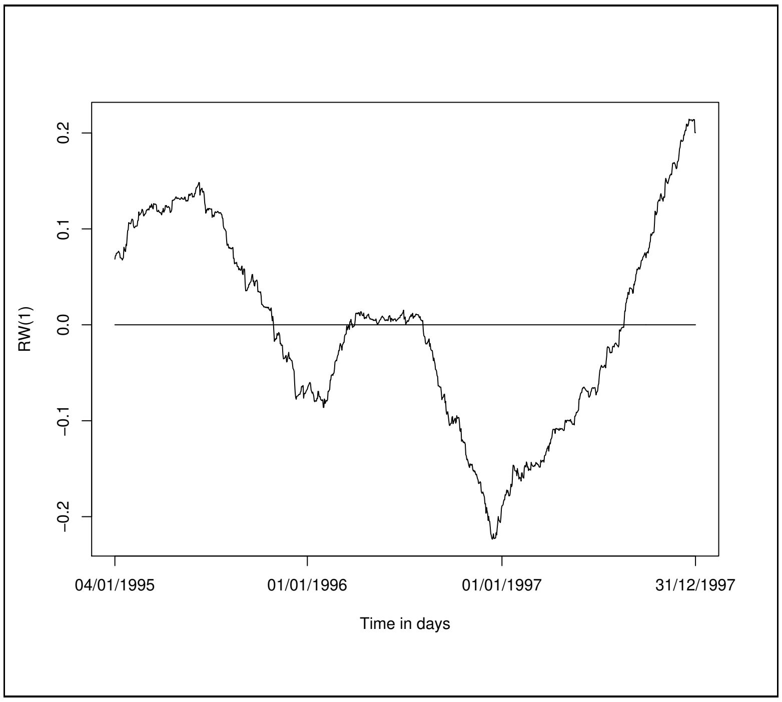

Figure 4: **Estimated effects of PM_{10} against time. The solid line is a smoothing spline of the median effects over time, the dashed line is the overall median of the effects of PM_{10} over the study period. The light shading shows 95% credible intervals from the full Bayesian analysis, and the dark shading is a 95% confidence interval for the fitted spline assuming the simulated median effects of PM_{10} to be true.**