

Modelling multiple pollutants at mutiple sites: a case study in Bayesian hierarchical modelling using WinBUGS

Gavin Shaddick

Department of Mathematical Sciences,
University of Bath.

UBC, October 2008

Aims

- ▶ Investigate the spatial-temporal modelling of pollutants.
- ▶ Assess the contribution of different components of variability; spatial, temporal and random variability.
- ▶ Develop methodology to provide:
 - ▶ exposures (and measures of uncertainty) for use in mapping of environmental factors
 - ▶ studies investigating the health effects of pollution.
- ▶ Fit models and perform analyses in WinBUGS.

Overview

- ▶ Background
- ▶ Data
 - ▶ Pollutant dependence
 - ▶ Temporal dependence
 - ▶ Spatial dependence
 - ▶ Missing values
 - ▶ Measurement error
- ▶ Models
 - ▶ Single pollutant, single monitoring site
 - ▶ Single pollutant, multiple monitoring sites
 - ▶ Multiple pollutants, single monitoring site
 - ▶ Multiple pollutants, multiple monitoring sites
- ▶ Summary
- ▶ Examples of implementation

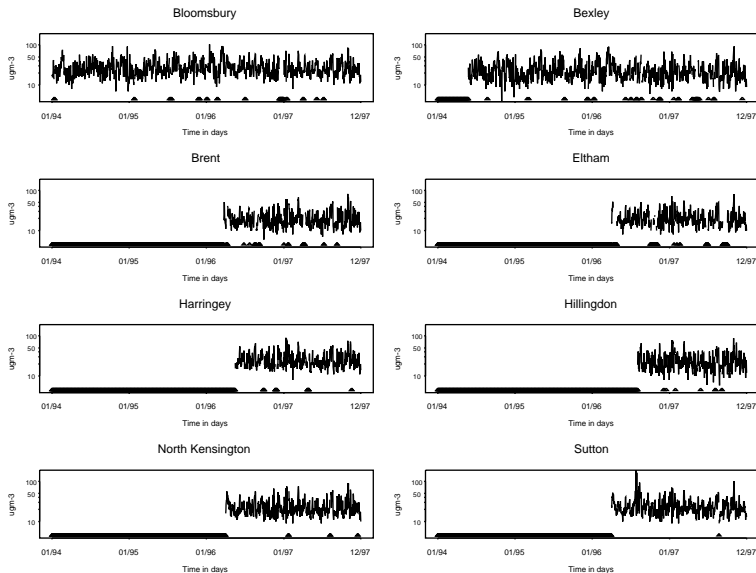
Background

- ▶ Daily measurements often available for different pollutants from a number of sites
- ▶ May be subject to measurement error
- ▶ Contain missing values
 - ▶ Pollutants not measured at all sites
 - ▶ Monitor being moved by design, e.g. six-day monitoring schedule
 - ▶ Unreliable or faulty monitors

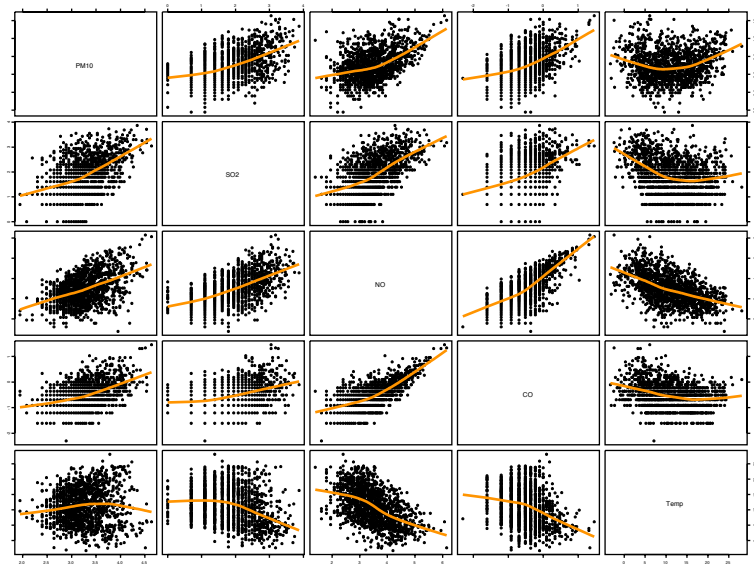
Data

- ▶ Eight sites within London, 1997-94
- ▶ PM₁₀, SO₂, NO and CO.
- ▶ All pollutants only measured at only 4 sites.
- ▶ Periods of operation between 1 and 4 years.
- ▶ Percentage of missing values as great as 37%.

Time series plots of (logged) values of PM₁₀



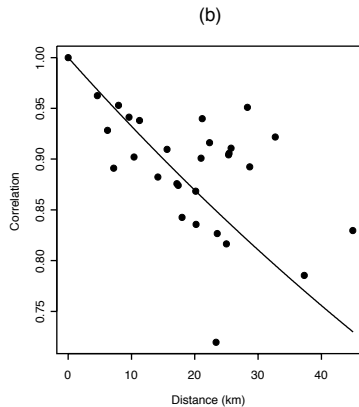
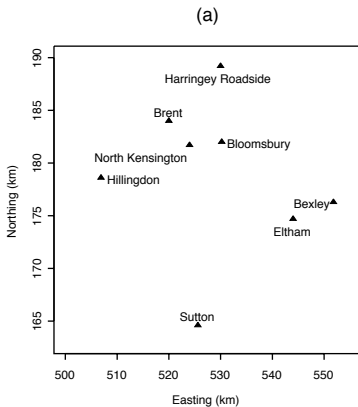
Correlations between pollutants and temperature



Data dependencies

- ▶ There are dependencies, both temporally and spatially, between daily measurements of different pollutants.
 - ▶ Pollutant dependence - common processes by which they are formed and the relationship with meteorological conditions.
 - ▶ Temporal dependence - atmospheric lifetimes and relationship with meteorological conditions.
 - ▶ Spatial dependencies - distance between sites and site type.

Locations of monitoring sites and correlations with distance



Model framework

- ▶ Bayesian hierarchical model.
- ▶ Pollutants modelled as a function of the true underlying level with measurement error.
- ▶ Incorporate covariate information, e.g. temperature.
- ▶ Underlying level is a function of the previous day's level.
- ▶ Missing values treated as unknown parameters within the Bayesian framework and can be estimated.

Single pollutant, single monitoring site

► Stage One, Observed Data Model:

$$Y_t = X_t^T \beta_1 + \theta_t + v_t,$$

v_t is referred to as *measurement error*, and assumed to be independent and identically distributed (i.i.d.) as $N(0, \sigma_v^2)$

► Stage Two, Temporal Model:


Autoregressive first order model

$$\theta_t = \rho \theta_{t-1} + w_t$$

w_t i.i.d. as $N(0, \sigma_w^2)$.

► Stage Three, Hyperprior:

Normal prior $N(c, C)$ for β_1 , where c is a $q_1 \times 1$ vector and C a $q_1 \times q_1$ variance-covariance matrix.

$\sigma_v^{-2} \sim Ga(a_v, b_v)$ and $\sigma_w^{-2} \sim Ga(a_w, b_w)$. 

Prior distribution
for $\theta' = (\theta_1, \dots, \theta_T)$

$$\begin{aligned} p(\theta|\sigma_w^2) &\propto \prod_{t=2}^T p(\theta_t|\theta_{t-1}, \sigma_w^2) \\ &\propto (\sigma^{-2})^{T-1} \exp \left\{ -\frac{1}{2\sigma_w^2} \sum_{t=2}^T (\theta_t - \theta_{t-1})^2 \right\} \\ &\propto (\sigma^{-2})^T \exp \left\{ -\frac{1}{2\sigma_w^2} \sum_{t=1}^T n_t \theta_t (\theta_t - \bar{\theta}_t) \right\} \end{aligned}$$

where n_t indicates the number of, and $\bar{\theta}$ the mean of, the neighbours of θ_t , i.e. θ_{t-1} and θ_{t+1} .

The prior distribution for θ , $p(\theta|\sigma_w^2)$, can therefore be expressed as

$$p(\theta_t|\theta_{-t}, \sigma_w^2) \sim \begin{cases} N(\theta_{t+1}, \sigma_w^2) & \text{for } t = 1, \\ N\left(\frac{\theta_{t-1} + \theta_{t+1}}{2}, \frac{\sigma_w^2}{2}\right) & \text{for } t = 2, \dots, T-1, \\ N(\theta_{t-1}, \sigma_w^2) & \text{for } t = T. \end{cases}$$

where θ_{-t} represents the vector of θ 's with θ_t removed.

Posterior distribution

The posterior distribution is given by

$$p(\theta, \beta_1, \sigma_v^2, \sigma_w^2 | y) = p(y)^{-1} \left\{ \prod_{t=1}^T p(y_t | \theta_t, \beta_1, \sigma_v^2) \right\} \times \\ \left\{ \prod_{t=2}^T p(\theta_t | \theta_{t-1}, \sigma_w^2) \right\} \times \\ p(\theta_1) p(\beta_1) p(\sigma_v^2) p(\sigma_w^2)$$

- ▶ Samples may be generated in a straightforward fashion using Markov chain Monte Carlo (using the WinBUGS software)
- ▶ Dealing with the cyclical graph that arises at stage two, requires some of the conditional distributions to be explicitly specified

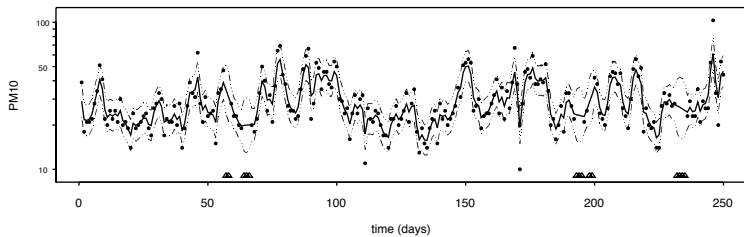
- ▶ Missing values are treated as parameters and the posterior obtained over these values and the model parameters. Samples can be generated from the distribution of missing values

$$p(y_m|y_o) = \int p(y_m|\lambda)p(\lambda|y_o)d\lambda$$

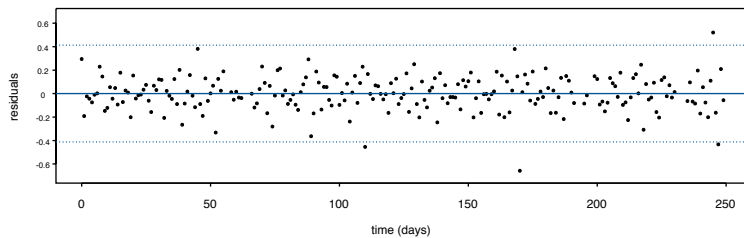
where $\lambda = (\theta, \beta_1, \sigma_v^2, \sigma_w^2)'$

Time series of 250 days of observed and estimated levels (together with their differences) of PM_{10} at
Bloomsbury

(a)



(b)



Single pollutant, multiple monitoring site

- ▶ S monitoring sites measuring a single pollutant.
- ▶ The underlying autoregressive structure remains constant across sites with a constant adjustment in the mean level for site s by an amount m_s , $s = 1, \dots, S$.
- ▶ **Stage One, Observed Data Model:**

$$Y_{st} = X'_{st}\beta_1 + X'_s\beta_2 + m_s + \theta_t + v_{st}$$

with v_{st} i.i.d. as $N(0, \sigma_{vs}^2)$ and β_1, β_2 , $q_1 \times 1$ and $q_2 \times 1$ vectors of site/day and site only regression coefficients.

- ▶ **Stage Two (a), Temporal Model:**

$$\theta_t = \theta_{t-1} + w_t$$

with w_t i.i.d. as $N(0, \sigma_w^2)$.

► **Stage Two (b), Spatial Model:**

The random effects $m = (m_1, \dots, m_S)'$ arise from the multivariate normal distribution

$$m \sim MVN(0_S, \sigma_m^2 \Sigma_m),$$

where 0_S is an $S \times 1$ vector of zeros,

σ_m^2 the between-site variance and

Σ_m is the $S \times S$ correlation matrix, in which element (s, s') represents the correlation between sites s and s' .

- This model is stationary and assumes an isotropic covariance model in which the correlation between sites s and s' is assumed to be a function of the distance between them

$$f(d_{ss'}, \phi) = \exp(-\phi d_{ss'})$$

where $\phi > 0$ describes the strength of the correlation

- A simpler model assumes that the site-specific levels are (conditionally) independent

$$m_s \sim \text{i.i.d } N(0, \sigma_m^2),$$

► Stage Three, Hyperpriors:

- Unless there is specific information to the contrary, i.e. that a monitor with different characteristics is used at a particular site, we will assume $\sigma_{vs}^{-2} \sim Ga(a_v, b_v)$.
- The between site precision has prior $\sigma_m^{-2} \sim Ga(a_m, b_m)$.
- A uniform prior is used for ϕ , with the limits being based on beliefs about the relationship between correlation and distance.
- The distance, d , at which the correlation, ρ , between two sites might be expected to fall to a particular level would be $d = -\log(\rho)/\phi$.

Estimating levels at unmeasured locations

- ▶ Based on the posterior estimates of the site effects, m_s and the variance-covariance matrix $\sigma_m^2 \Sigma_m$, it is possible to estimate the site effects, and thus pollution levels, at locations where there is no monitoring site.
- ▶ For a site at a new location, m_{S+1} , $(m_1, \dots, m_S, m_{S+1})$ follows a multivariate normal distribution with zero mean and $(S + 1) \times (S + 1)$ variance-covariance matrix.
- ▶ Letting $m = (m_1, \dots, m_S)'$, the conditional distribution of $m_{S+1}|m$ is, normal with mean and variance given by

$$E[m_{S+1}|m] = \sigma_m^{-2} \Omega' \Sigma_m^{-1} m,$$

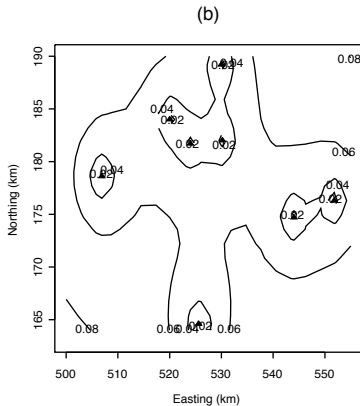
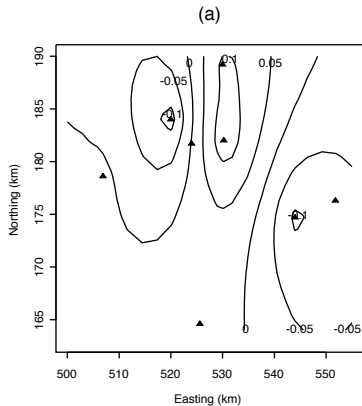
$$\text{var}(m_{S+1}|m) = \sigma_m^2 (1 - \Omega' \Sigma_m^{-1} \Omega),$$

- ▶ For exploratory purposes, the posterior medians may be substituted into these expressions (although this will ignore the inherent uncertainty in the estimates).

Site effects

	Median	2.5%	97.5%
Bexley	-0.0696	-0.0785	-0.0607
Bloomsbury	0.1341	0.1257	0.1426
Brent	-0.1210	-0.1294	-0.1125
Eltham	-0.1105	-0.1205	-0.1005
Harringey	0.1098	0.0999	0.1195
Hillingdon	0.0132	-0.0032	0.0300
North Kensington	0.0030	-0.0031	0.0090
Sutton	0.0410	0.0250	0.0572
σ_m	0.1019	0.0668	0.1794
ϕ	0.05675	0.02158	0.09778

Contour plot of site effects based on a 20x20 grid of locations without a pollution monitor with corresponding standard deviations



Multiple pollutants, single monitoring site

► Stage One, Observed Data Model:

$$Y_{pt} = X_t' \beta_1 + \theta_{pt} + v_{pt}$$

with v_{pt} i.i.d. as $N(0, \sigma_{vp}^2)$ and β_1 a $q_1 \times 1$ vector of regression coefficients.

► Stage Two, Temporal and Pollutant Model:

$$\theta_{pt} = \theta_{p,t-1} + w_{pt}$$

$w_t = (w_{1t}, \dots, w_{Pt})'$ are i.i.d. multivariate normal random variables with zero mean and variance-covariance matrix Σ_P .

► Stage Three, Hyperpriors:

$$\sigma_{vp}^{-2} \sim Ga(a_v, b_v), \quad p = 1, \dots, P.$$

$\Sigma_P^{-1} \sim W_P(D, d)$, a P -dimensional Wishart distribution with mean D and precision parameter d .

- ▶ Model was applied to data from four pollutants (PM_{10} , SO_2 , NO and CO) from the Bloomsbury site.
- ▶ Priors $\sigma_{vp}^{-2} \sim Ga(1, 0.01)$, $p = 1, \dots, P$, and $\beta_1 \sim N(0, 1000)$.
- ▶ For the parameters of the Wishart distribution, d was chosen to be equal to four, the dimension of Σ_P ;
 D was then chosen so that the diagonals of the expected value (D/d) represent a 10% coefficient of variation. The off-diagonals were taken to be zero.
- ▶ Posterior correlations

	PM_{10}	SO_2	NO	CO
PM_{10}	1.0000	0.8806	0.8192	0.8134
SO_2	0.8806	1.0000	0.8472	0.9202
NO	0.8192	0.8472	1.0000	0.9146
CO	0.8134	0.9202	0.9146	1.0000

- ▶ Strong correlations mean that inference on missing values can be made on the values of pollutants

Multiple pollutants, multiple monitoring sites

► Stage One, Observed Data Model:

$$Y_{spt} = X'_{pt}\beta_1 + X'_{st}\beta_2 + \theta_{pt} + m_s + v_{spt},$$

where v_{spt} are i.i.d. $N(0, \sigma_{sp}^2)$, β_1 a $q_1 \times 1$ vector of pollutant regression coefficients, and β_2 a $q_2 \times 1$ vector of spatial regression coefficients.

► Stage Two, Spatial, Temporal and Pollutant Model:

The $(p \times 1)$ vector of daily pollution measurements, $(\theta_1, \dots, \theta_P)'$, as a function of the previous days values with possible correlation between the values of the different pollutants.

An alternative approach would be to allow the spatial effects to be pollutant specific

► Stage Three, Hyperprior:

In the absence of additional information, we assume that $\sigma_{vsp}^{-2} \sim Ga(a_v, b_v)$.

Components of variability

- ▶ Model 1 (Single pollutant, single site)
 - ▶ Temporal 70%
 - ▶ Measurement error 30%
- ▶ Model 2 (Single pollutant, multiple sites)
 - ▶ Temporal 80%
 - ▶ Spatial 10%
 - ▶ Measurement error 10%
- ▶ Model 3 (Multiple pollutants, single site)
 - ▶ Temporal 77%
 - ▶ Measurement error 23%
- ▶ Model 4 (Multiple pollutants, multiple sites)
 - ▶ Temporal 75%
 - ▶ Spatial 15%
 - ▶ Measurement error 10%

Summary

- ▶ Examine the contribution of spatial, temporal and random variability.
- ▶ Measures of uncertainty, with implications on the precision of the resulting relative risks.
- ▶ Allows levels to be estimated at non-measured locations.
- ▶ Calculate underlying levels of pollution for use in health studies.
- ▶ Estimates of missing values.

The assumptions of the model include the following:

- ▶ The measurement error variance σ_{sp}^2 does not depend on time. The model is easily extendable to situations in which the measurement error may change as a function of t , for example, when a monitor is replaced.
- ▶ The relationship between the pollutants is constant over time.
- ▶ The relationship between the pollutants is spatially constant.
- ▶ The temporal and spatial components are independent.

Examples of implementation of the model framework

- ▶ Spatial-temporal model - using modelled levels of PM_{10} in a health study.
- ▶ Spatial model - mapping concentrations of SO_2 over entire EU.

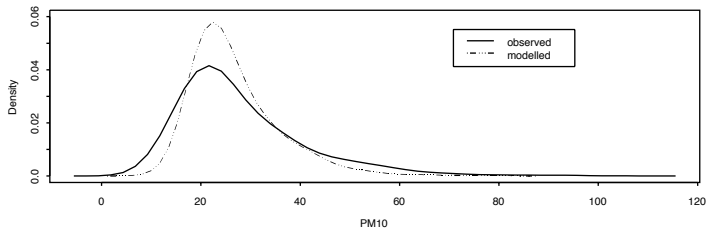
Health analysis

- ▶ PM_{10} and respiratory mortality (ICD 460-519) in London, 1994-97.
- ▶ Assess the effects of using modelled levels of pollutant on relative risks.
- ▶ Base model contains terms for trend, trend², year, month, year \times month interaction, day of week, 12, 6, 4 and 2 monthly cycles and temperature (same day, lag 1, lag2).

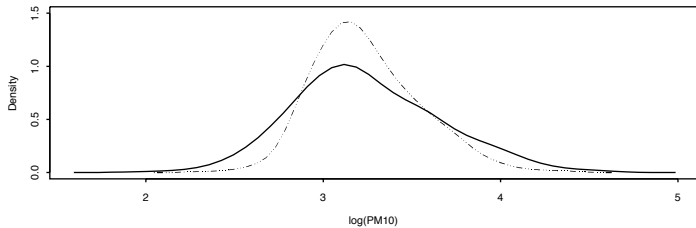
Distributions of observed and modelled values of PM_{10}

(normal and logged values)

(a)



(b)



Relative risks (and 95% CIs) associated with increase of $10\mu\text{gm}^{-3}$ in PM_{10} (lag 1)

- ▶ Observed PM_{10} with missing values excluded
 - ▶ $\text{RR} = 1.0116$ (1.0046 - 1.0186)
- ▶ Modelled PM_{10} with missing values excluded
 - ▶ $\text{RR} = 1.0166$ (1.0064 - 1.0269)
- ▶ Modelled PM_{10} with estimated missing values
 - ▶ $\text{RR} = 1.0182$ (1.0084 - 1.0280)

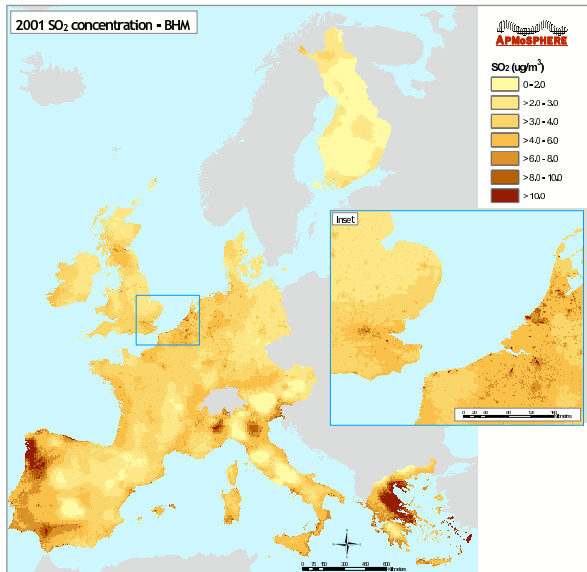
using spatial model

- ▶ Modelled PM_{10} with missing values excluded
 - ▶ $\text{RR} = 1.0134$ (1.0066 - 1.0203)
- ▶ Modelled PM_{10} with estimated missing values
 - ▶ $\text{RR} = 1.0128$ (1.0062 - 1.0195)

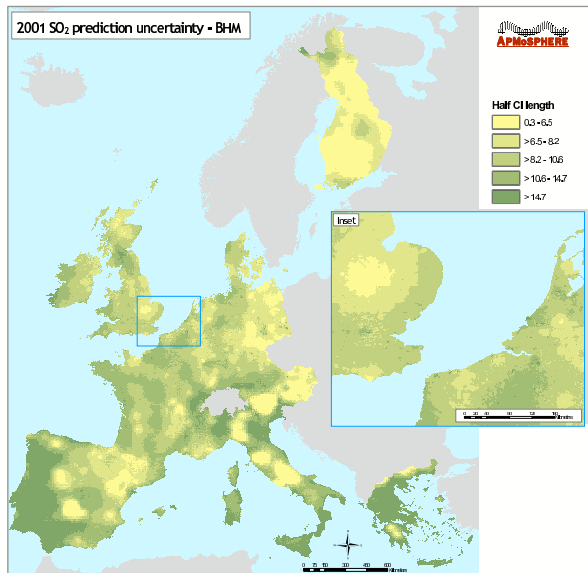
Large scale mapping of SO₂ over the entire EU.

- ▶ This used data from the APMoSPHERE project (www.apmosphere.org).
- ▶ Concentrations of SO₂ were obtained from 253 monitoring stations located non-uniformly over the EU.
- ▶ High resolution (at the 1km × 1km level) climatic and geographical information was also obtained, including seasonal value rainfall and temperature, wind speed, altitude and distance to sea.
- ▶ Due to the high levels of collinearity observed in the climate variables, principal component analysis (PCA) was used to reduce the original nine variables to five factors, which accounted for 97% of the total variation.

Predicted concentrations of SO₂ using Bayesian Hierarchical model.



Length prediction of 95% credible intervals.



Implementing the temporal model in WinBUGS