



# Bayesian model averaging approach in health effects studies: Sensitivity analyses using PM<sub>10</sub> and cardiopulmonary hospital admissions in Allegheny County, Pennsylvania and simulated data

Ya-Hsiu Chuang<sup>1</sup>, Sati Mazumdar<sup>1</sup>, Taeyoung Park<sup>2</sup>, Gong Tang<sup>1</sup>, Vincent C. Arena<sup>1</sup>, Mark J. Nicolich<sup>3</sup>

<sup>1</sup> Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

<sup>2</sup> Department of Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. Presently at Dept. of Applied Statistics, Yonsei University, Seoul, Korea

<sup>3</sup> Lambertville, New Jersey and previously with ExxonMobil Biomedical Sciences, Inc., Annadale, New Jersey, USA

## ABSTRACT

Generalized Additive Models (GAMs) with natural cubic splines (NS) as smoothing functions have become a standard analytical tool in time series studies of health effects of air pollution. However, standard model selection procedures ignore the model uncertainty that may lead to biased estimates, in particular those of the lagged effects. We addressed this issue by Bayesian model averaging (BMA) approach which accounts for model uncertainty by combining information from all possible models where GAMs and NS were used. Firstly, we conducted a sensitivity analysis with simulation studies for Bayesian model averaging with different calibrated hyperparameters contained in the posterior model probabilities. Our results indicated the importance of selecting the optimum degree of lagging for variables, based not only on maximizing the likelihood, but also by considering the possible effects of concurvity, consistency of degree of lagging, and biological plausibility. This was illustrated by analyses of the Allegheny County Air Pollution Study (ACAPS) where the quantity of interest was the relative risk of cardiopulmonary hospital admissions for a 20 µg/m<sup>3</sup> increase in PM<sub>10</sub> values for the current day. Results showed that the posterior means of the relative risk and 95% posterior probability intervals were close to each other under different choices of the prior distributions. Simulation results were consistent with these findings. It was also found that using lag variables in the model when there is only same day effect, may underestimate the relative risk attributed to the same day effect.

### Keywords:

Health effects  
Lagged effects  
Bayesian model averaging  
Simulation studies

### Article History:

Received: 20 April 2010

Revised: 01 July 2010

Accepted: 11 July 2010

### Corresponding Author:

Vincent C. Arena

Tel: +1-412-6243023

Fax: +1-412-6242183

E-mail: arena@pitt.edu

© Author(s) 2010. This work is distributed under the Creative Commons Attribution 3.0 License.

doi: 10.5094/APR.2010.021

## 1. Introduction

Generalized additive models (GAMs) have been used as a standard analytical tool to investigate the effect of air pollution on public health in time series studies. Due to the characteristics of time series data, the effects of long-term trends and seasonality, meteorological variability, and day of the week effects need to be removed. GAMs have the advantage of allowing non-linear relationships between predictor variables and the selected response. The smoothers and the degrees of smoothing for the predictor variables need to be specified in the fit of GAMs. The most common choices for smoothers are natural cubic spline, smoothing spline, and LOESS, where natural cubic spline is a parametric smoother, and smoothing spline and LOESS are the nonparametric smoothers. When a natural cubic spline is used in a GAM, it becomes a fully parametric generalized linear model (GLM). The model building procedures for both GAMs and GLMs usually follow the standard rule, where a subset of predictor variables gets selected according to their statistical significance levels. However, as the predictor variables are usually found to be multicollinear, selection of these variables becomes a major statistical issue.

Let us consider the issue of the lagged effects of ambient air levels of a criteria pollutant (e.g. PM<sub>10</sub>: particulate matter with diameter 10 µm or less) on cardiopulmonary distress. Theoretically, the effect of PM<sub>10</sub> on cardiopulmonary distress can last for more than one day. Therefore, it is important to find exactly how

long this effect usually lasts. Using data from the study in Birmingham, Smith et al. (2000) applied standard model selection procedures to determine the number of lag variables of different lengths for PM<sub>10</sub> and found that none of the lag variables were statistically significantly associated with non-accidental elderly mortality. However, Schwartz (1993) used the average of PM<sub>10</sub> for the three previous days and found a statistically significant effect between PM<sub>10</sub> and non-accidental elderly mortality. In the analysis of Allegheny County Air Pollution Study (ACAPS) where the health effects of PM<sub>10</sub>, on daily hospital admissions from cardiopulmonary disease in Allegheny County, Pittsburgh, PA from 1995 through 2000 was assessed and only the same day level of PM<sub>10</sub> was found to have an effect (relative risk 1.012256). Wordley et al. (1997) used Birmingham, UK, data from 1992 to 1994 and included PM<sub>10</sub> on the same day, lagged by up to three days, and a three day mean (mean of the same day and the two previous days) as the effect of PM<sub>10</sub> in the model. Statistically significant associations of these variables with all respiratory hospital admissions were found. However, the standard model selection approaches (e.g., AIC criterion was used in the ACAPS) did not take into account uncertainties associated with them.

Bayesian model averaging (BMA) provides an approach to take into account model uncertainty by combining information from a pre-determined subset of all possible models and obtaining a weighted average for the quantity of interest over these models (Hoeting et al., 1999). One advantage of BMA is that it can include all predictor variables in the model. Variables that are less

important have smaller weights. Implementation of BMA requires the specification of prior distributions for parameters within models and prior weights for each model. Clyde (2000) developed a class of objective prior distributions for parameters within models. These objective prior distributions have a hyperparameter that is used to calibrate the priors based on classical model selection criteria. As the conclusions can be sensitive to the choices of the hyperparameter, Clyde (2000) suggested providing estimates for several prior distributions, i.e., from several choices of the hyperparameter, thus suggesting a sensitivity analysis (Clyde, 2000). Applications of Bayesian methods have recently been seen in air pollution studies (Nikolov et al., 2007; Lee and Shaddick, 2008; Liu et al., 2008).

Section 2 presents a brief description of BMA and methods for its implementation. An illustrative example using ACAPS data is presented in Section 3. A simulation study is given in Section 4 followed by a discussion in Section 5.

As background to this paper we provide a brief summary of ACAPS where the health effects of PM<sub>10</sub>, on daily hospital admissions from cardiopulmonary disease was assessed in Allegheny County from 1995 through 2000 (Arena et al., 2006). They derived models of daily hospital admissions from cardiopulmonary disease as a function of daily mean level of PM<sub>10</sub>, weather, long term trends and seasonality and day of the week using generalized additive models (GAM) with locally weighted regression smoother (LOESS). Models were derived using same day as well as up to five previous days of PM<sub>10</sub> levels. Findings suggested that there is a positive association of current day PM<sub>10</sub> levels with cardiopulmonary hospital admissions in this population independent of long-term trends and seasonality, weather (average daily temperature and average daily relative humidity), and the day of the week. Considering a 20 µg/m<sup>3</sup> change in current day PM<sub>10</sub>, the estimate of the relative risk was 1.012256.

**2. Bayesian Model Averaging**

**2.1. Bayesian model averaging (BMA)**

BMA starts with a set of plausible models and averages the posterior distributions of the quantity of interest obtained under each of these models, weighted by the corresponding posterior model probabilities. Let  $\Lambda$  denote the quantity of interest that has the same interpretation in each of the models considered (e.g. the relative risk associated with a particular increment in the air pollutant level on health outcome). The posterior distribution of  $\Lambda|Y$  can be written as:

$$\Pr(\Lambda | Y) = \sum_{m=1}^K \Pr(\Lambda_m | Y, M_m) \Pr(M_m | Y) \tag{1}$$

where  $M_m$  is the  $m^{th}$  model under consideration and  $\Lambda_m$  is the quantity of interest in  $M_m$  with  $m=1, \dots, K$  and  $K$  is the size of the set of all models being considered. The first term on the right hand side of Equation (1) is the posterior distribution of  $\Lambda_m$  given a particular model  $M_m$  and the data, and the second term is the posterior probability of the model  $M_m$ .

**2.2. Implementation of BMA**

The posterior distribution of  $\Lambda$  given a particular model  $M_m$  and data  $Y$  in Equation (1) is given by:

$$\Pr(\Lambda_m | Y, M_m) = \int \Pr(\Lambda_m | \beta_m, M_m, Y) \Pr(\beta_m | M_m, Y) d\beta_m \tag{2}$$

where,  $\beta_m$  is the vector of parameters for the model  $M_m$ . As Equation (2) may not provide any closed form solutions, we used maximum likelihood estimate (MLE) of  $\beta_m$  to approximate it giving:

$$\Pr(\Lambda_m | Y, M_m) \approx \Pr(\Lambda_m | \hat{\beta}_m, Y, M_m) \tag{3}$$

The posterior probability for model  $M_m$  is given by:

$$\Pr(M_m | Y) = \frac{\Pr(Y | M_m) \Pr(M_m)}{\sum_{j=1}^K \Pr(Y | M_j) \Pr(M_j)} \tag{4}$$

where,

$$\Pr(Y | M_m) = \int \Pr(Y | \beta_m, M_m) \Pr(\beta_m | M_m) d\beta_m \tag{5}$$

$\Pr(\beta_m | M_m)$  is the prior density of  $\beta_m$  under model  $M_m$ , and  $\Pr(M_m)$  is the prior density of the model  $M_m$ . In order to derive the posterior model probability, these prior densities for models and parameters within each model need to be specified in advance.

We followed the formulation of Clyde (2000) for a generalized linear model. The prior distributions for the regression parameters in the models that describe the relationship between the outcome variable and the explanatory variables were objective priors based on Jeffrey’s modification of Calibrated Information Criterion (CIC) prior distributions with a hyperparameter  $g$ . Specific choices of  $g$  reconciled classical model selection with Bayesian model selection based on posterior model probabilities. For the calibration of posterior model probabilities, we used uniform priors, i.e., non-informative priors on different models ( $M_m, m=1, 2, 3, \dots, K$ ).

Thus we have  $\Pr(M_m) = \pi(M_m) \sim$  uniform and Jeffrey’s modification of the CIC prior distribution under model  $M_m$

$$\Pr(\beta_m | M_m) \pi(M_m) = (2\pi)^{-d_m/2} \left| \frac{I(\hat{\beta}_m)}{g} \right|^{1/2} \prod_{j=1}^p \delta_0(\beta_j)^{1-\gamma_j} \tag{6}$$

where  $d_m$  is the dimension of model  $M_m$ ,  $g$  is the hyperparameter used in calibration of posterior model probabilities,  $I(\hat{\beta}_m)$  is the observed Fisher information for  $M_m$  evaluated at the MLEs  $\hat{\beta}_m$  with  $(j,k)th$  elements,

$$[I(\beta_m)]_{jk} = - \left[ \frac{\partial^2}{\partial \beta_j \partial \beta_k} L(\beta | M_m) \right] \text{ with } L(\beta | M_m) \text{ as the log likelihood}$$

under  $M_m$ ,  $\gamma_j$  is the indicator variable, 1 if  $x_j$  is included under  $M_m$ , 0 otherwise,  $\delta_0(\beta_j)$  is the degenerate distribution that degenerates at 0 if variables are not in  $M_m$ .

For the Poisson regression model with log link, the observed information matrix is  $I(\beta_m) = X_m^T V(\beta_m) X_m$ , where  $V(\beta_m)$  denotes the covariance matrix for  $Y$  with elements  $\exp(X_m \beta_m)$  on the diagonal and zero elsewhere. The posterior model probability is then given by

$$\Pr(M_m | Y) = \frac{\exp[0.5 (D_m - d_m \log(g))]}{\sum_{j=1}^K \exp[0.5 (D_j - d_j \log(g))]} \tag{7}$$

where  $D_m$  is the model deviance which is the usual deviance (-2 times the log likelihood) under the null model minus the deviance under  $M_m$ ,  $d_m$  is the dimension of  $\beta_m$ , and  $g$  is the hyperparameter (Clyde, 2000).

A second issue for the implementation of BMA is to find data-supported models. There are up to  $2^p$  possible models when  $p$  predictor variables are under consideration. As  $p$  increases, the number of models in BMA becomes larger leading to computationally expensive operations. Moreover, many of these models may have very little support from the data and their

inclusion will not have practical importance. One way to approximate Equation (1) is by averaging over the better models only. Madigan and Raftery (1994) proposed Occam’s Window approach that includes models with the higher posterior model probabilities and excludes models with posterior model probabilities lower than any of their simpler sub-models. The posterior mean and variance of  $\Lambda$  are given by Hoeting et al. (1999):

$$E(\Lambda | Y) = \sum_{m=1}^K E(\Lambda_m | Y, M_m) \Pr(M_m | Y) \tag{8}$$

$$Var(\Lambda | Y) = \sum_{m=1}^K ((Var(\Lambda_m | Y, M_m) + (E(\Lambda_m | Y, M_m))^2) \cdot \Pr(M_m | Y) - E(\Lambda | Y)^2) \tag{9}$$

In the CIC  $g$ -prior of the parameters  $\theta_m$  the choice of  $g$  controls model selection in a way that small  $g$  tends to concentrate the prior on saturated models with small coefficients and large  $g$  concentrates the prior on parsimonious models with a few large coefficients (George and Foster, 2000). It has been shown that the posterior model probabilities under a  $g$ -prior can be calibrated to different classical model selection criteria such as AIC and BIC (Clyde, 2000). In addition, the Empirical Bayes (EB) approach was developed to provide adaptive estimates of  $g$ . The local EB approach (George and Foster, 2000; Hansen and Yu, 2003; Hansen and Yu, 2001) estimates  $g$  from the data and assumes that different models have different estimates of  $g$ .

In this paper, we have implemented BMA under:

(i) AIC prior, where the posterior model probabilities under this prior can be calibrated to the classical model selection criterion in AIC by using  $\log(g) = 2$ ;

(ii) BIC prior, where the posterior model probabilities under this prior can be calibrated to the classical model selection criterion in BIC by using  $\log(g) = \log(n)$  with  $n$  as the number of observations; and

(iii) local EB approach estimate of  $\hat{g}_m^{EBL}$ , where  $\hat{g}_m^{EBL}$  is the MLE for  $g$  by using the local EB approach and is constrained to be nonnegative. This estimate of  $g$  is given by

$$\hat{g}_m^{EBL} = \max\left(\frac{\hat{\beta}_m^T I(\hat{\beta}_m) \hat{\beta}_m}{d_m} - 1, 0\right) \text{ for a GLM with dispersion parameter}$$

of 1, where  $\hat{\beta}_m$  is the MLE of  $\beta_m$  and  $d_m$  is the dimension of model  $M_m$ .

The BMA approach was implemented by modifying the S-Plus program that calculates the BMA based on BIC, *bic.glm*, to correspond to the prior choices based on AIC, BIC, and local EB approach.

The quantity of interest in this paper is the relative risk associated with air pollutant level on cardiopulmonary hospital admissions. We used the following formulas to calculate it. Based on a  $20 \mu\text{g}/\text{m}^3$  increase in all the  $\text{PM}_{10}$  variables ( $\text{PM}_{10\_lag0}, \dots, \text{PM}_{10\_lagm}$ ), in model  $M_m$  the relative risks for each model were given by:

$$\Lambda_m = \exp \left[ 20 \left( \beta_{\text{PM}_{10\_lag0}} + \beta_{\text{PM}_{10\_lag1}} + \dots + \beta_{\text{PM}_{10\_lagm}} \right) \right] \tag{10}$$

where  $m$  is the lag length of the  $\text{PM}_{10}$ .

The posterior distribution for the relative risk given  $M_m$  follows a log-normal distribution

$$\log(\Lambda_m) | M_m, Y \sim N \left[ 20 \left( \beta_{\text{PM}_{10\_lag0}} + \dots + \beta_{\text{PM}_{10\_lagm}} \right), \sigma_{\Lambda_m}^2 \right] \tag{11}$$

where  $\sigma_m^2 = 20^2 (1^T \Sigma_{\beta_m | M_m} 1)$  with  $1^T = (1, \dots, 1)$  of dimension  $m$  and  $\Sigma_{\beta_m | M_m}$  is the covariance matrix for the  $\text{PM}_{10}$  variables under model  $M_m$  derived from the Fisher information under model  $M_m$  (Clyde, 2000). The posterior means of the log-relative risk and the 95% posterior probability intervals are calculated using Equations (8), (9) and (11).

### 3. Application of BMA method to the ACAPS data

#### 3.1. Starting model for ACAPS data

ACAPS contained time series data for the counts of daily cardiopulmonary hospital admissions, daily meteorological data, and daily ambient air levels of a criteria pollutant ( $\text{PM}_{10}$ ) for Allegheny County from 1995 to 2000 (Arena et al., 2006). The daily cardiopulmonary hospital admissions included records with a discharge diagnosis of the circulatory system or respiratory system for Allegheny County residents >65 years of age. The daily mean temperatures were used as the meteorological data in our study. Ambient air levels of a criteria pollutant ( $\text{PM}_{10}$ ) were recorded in every hour for each of the 8 monitoring sites. The mean of the site-specific daily average  $\text{PM}_{10}$  values across all monitoring sites was used as the pollutant level. Since only two sets of data out of 2 192 were missing on dates 03/24/1998 and 11/04/1998, they were ignored and we had a total of 2 190 observations for data analysis.

Arena et al. (2006) used unconstrained lag models to evaluate the association between  $\text{PM}_{10}$  and daily cardiopulmonary hospital admissions. They used GAM to fit the logarithm of the number of daily hospital admissions as a sum of smooth functions of long-term trends and seasonality, temperature and relative humidity, day of the week, and  $\text{PM}_{10}$  levels for the current day and previous days up to five days. LOESS was used as the smooth function with smoothing parameters (span) of 0.06 for seasonal trends and 0.5 for temperature and relative humidity. AIC criterion was used for model selection.

The humidity included in ACAPS study did not show significant association with the hospital admissions. Therefore, we did not include it as a predictor variable in the present study. Hence, the selected predictor variables in the present study included the levels of  $\text{PM}_{10}$  for same day and lagged up to five days ( $\text{PM}_{10\_lag0}, \dots, \text{PM}_{10\_lag5}$ ), the daily mean temperature (temp), the seasonal trend (time), and day of the week (DOW), which consists of six indicator variables. The natural cubic spline was used as the smooth function. We based this choice based on the following. When considering GAMs with smoothing spline and GLMs with natural cubic spline, He et al. (2006) showed that GLM with natural cubic spline performs better with respect to the bias and variance estimates when concurvity exists in the data. Concurvity is a nonparametric analogue of multicollinearity where a function of a predictor can be approximated by a linear combination of functions of other predictors (Ramsay et al., 2003).

In our ACAPS, the degrees of smoothing for the long-term trend and seasonality were determined by fitting the smooth function of long-term trend and seasonality with a range of degrees of smoothing on cardiopulmonary hospital admissions using GLMs with natural cubic splines. They were chosen from the fitted model that has the smallest AIC; the smaller AIC indicating the better the model fit. In addition, the residual plots were used to examine whether the seasonal variation has been removed. We then considered the short-term effects by adding six indicator variables for day of the week and the smooth function of temperature into the model and repeated the same procedure to find the degrees of smoothing for the temperature variable. This resulted in 5 degrees of freedom per year for long-term trend and

seasonality, and 7 degrees of freedom for daily mean temperature. We note that all subsequent analyses are conditional to this starting model.

The GLM with natural cubic spline used in this paper is given by:

$$Y_t \sim \text{Poisson}(\mu_t)$$

$$\log(\mu_t) = \log(\mu_0) + \beta_0 PM_{10\_lag0} + \dots + \beta_5 PM_{10\_lag5} + ns(\text{time}, df = 5 / \text{year}) + ns(\text{temp}, df = 7) + \eta I_{DOW} \quad (12)$$

where  $Y_t$  is the counts of daily cardiopulmonary hospital admissions, which we assume to follow a Poisson distribution (count data) with mean  $\mu_t$  and dispersion parameter of 1,  $PM_{10\_lag0}, \dots, PM_{10\_lag5}$  are the levels of  $PM_{10}$  for same day and lagged up to five days,  $ns(\text{time}, df = 5/\text{year})$  is the natural cubic spline function of calendar time with 5 degrees of freedom per year,  $ns(\text{temp}, df = 7)$  is the natural cubic spline function of temperature with 7 degrees of freedom,  $I_{DOW}$  are the six indicator variables for days of the week.

### 3.2. Bayesian model averaging analysis

The model given in Equation (12) includes 49 predictor variables resulting in  $2^{49}$  possible models.

Occam's Window was applied to find the data-supported models through the modified bic.glm package in S-Plus (<http://www2.research.att.com/~volinsky/software/bic.glm>). It ranked the models and we used the 150 top ranked models that had the highest posterior model probabilities. To examine which predictor variables were chosen under each of the selected models, we constructed model matrices for BMA under the three priors. Model matrices have the advantages of not only allowing us to visually identify which variables have more influence on the outcome variable but also reflect model uncertainties through the posterior model probabilities. The top 25 models under AIC, BIC, and local EB estimate are shown in Figure 1.

Figure 1 also includes posterior model probabilities and probabilities of inclusion of the predictor variables in the models. The y-axis represents the selected models ordered from the best to the worst (moving from bottom to top) based on the ranks using posterior model probabilities that are given on the right side. The x-axis shows the predictor variables included in the model, the predictor variables are given on the top of the figures and the probabilities of inclusion in the model are given in the bottom of the figures. The names of the predictor variables with "time" and "temp" on the x-axis represent the smooth functions for long-term trend and seasonality and for daily mean temperature, respectively. The number followed by "time" or "temp" is the knot number specified through the degree of freedom of the natural cubic spline. The dark squares in the matrix represent the predictor variables that were excluded under a given model. The histograms in Figure 2 show the posterior distributions of the relative risk for an increase of  $20 \mu\text{g}/\text{m}^3$  in the same day level for the three prior choices. Comparing to BIC prior, posterior distributions under AIC prior and EB estimate were found to be more dispersed indicating more uncertainties of model and parameters. The green lines in the histograms represent the posterior means.

The posterior means of the relative risk and the 95% posterior probability intervals derived from Equations (8), (9) and (11) were reported in Table 1. Based on a  $20 \mu\text{g}/\text{m}^3$  increase in the same day  $PM_{10}$  variables ( $PM_{10\_lag0}$ ), the posterior means of the relative risk ranged 0.9980 to 1.0022. The posterior probability intervals for BMA with the BIC prior and local EB estimate were found wider

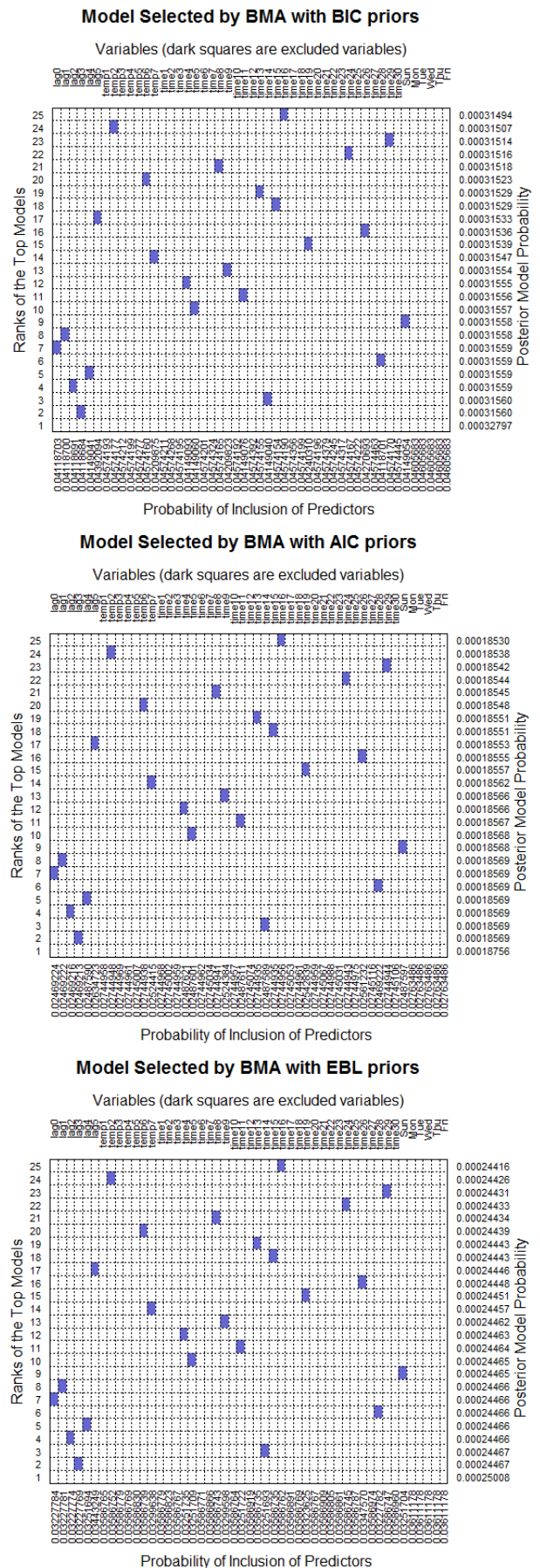
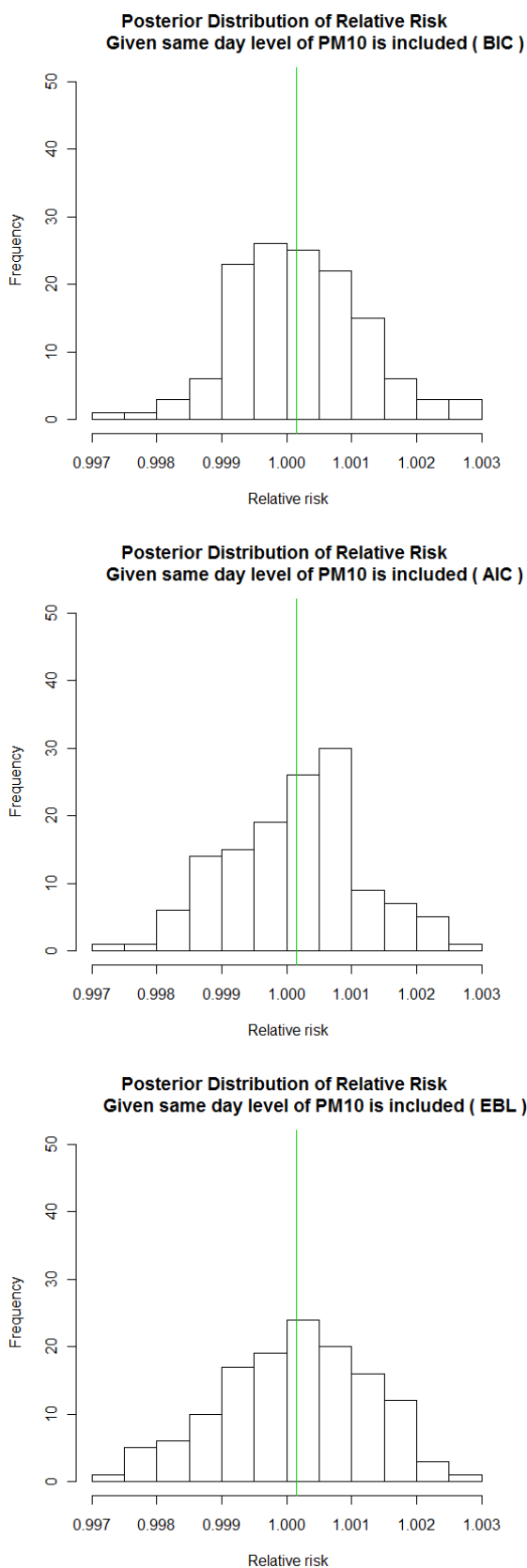


Figure 1. Plots of model space (BIC, AIC and EBL).

than those for BMA with the AIC prior. BIC prior and local EB estimate utilize the information from the data to estimate the hyperparameter  $g$  and this could lead to some greater levels of uncertainty.

**Table 1.** Summary of the posterior distribution of relative risk associated with a 20  $\mu\text{g}/\text{m}^3$  increase in the same day level of  $\text{PM}_{10}$  under BMA

Prior	Posterior mean of relative risk	95% posterior probability interval of relative risk
AIC	1.0001	(0.9984, 1.0017)
BIC	1.0001	(0.9980, 1.0022)
Local EB	1.0001	(0.9982, 1.0020)



**Figure 2.** Posterior distributions of relative risks given same day level of  $\text{PM}_{10}$  included (BIC, AIC and EBL).

We should note here that our approach of choosing the predictor variables during the BMA procedure is novel. We started with a model with spline smoothers with degrees of freedom 5/year for time and 7 for temp. Under a Bayesian perspective, it has been increasingly common and standard to use splines for which the number and location of knots are free parameters (DiMatteo et al., 2001; Holmes and Mallick, 2003). Because we consider including/excluding each basis function separately, our approach for smooth functions can be viewed as the method of free-knot splines, which is more flexible and parsimonious than treating each smooth function as a single unit for inclusion/exclusion. The free-knot splines are also advantageous because the amount of data smoothing can be determined in a locally adaptive manner by including/excluding each basis function separately.

**4. Simulation Study**

To demonstrate how the results from BMA approach under different prior choices vary, we provided a simulation study. Following the earlier work of simulation procedures in He et al. (2006), we generated the time series data using ACAPS data.

To generate a 6-year hospital admissions time series, we used the following model:

$$Y_t \sim \text{Poisson}(\mu_t)$$

$$\log(\mu_t) = \log(\mu_0) + \beta_0 \widetilde{PM}_{10,lag0} + 0.25 \text{ time} + \text{temp. } s + \eta I_{DOW} \quad (13)$$

$\mu_0$  in Equation (13) represents the mean of daily cardiopulmonary hospital admissions over the 6-year period and was estimated from ACAPS data as 115.07.  $\beta_0$  is the true  $\text{PM}_{10}$  effect and  $\eta$  are the true effects for day of the week. Both effects were initially estimated by fitting the following model to the observed ACAPS data:

$$Y_t \sim \text{Poisson}(\mu_t)$$

$$\log(\mu_t) = \log(\mu_0) + \beta_0 PM_{10,lag0} + ns(\text{time}, df = 5/\text{year}) + ns(\text{temp}, df = 7) + \eta I_{DOW} \quad (14)$$

where  $Y_t$  is the counts of daily cardiopulmonary hospital admissions, which follows a Poisson distribution with mean  $\mu_t$ ,  $\beta_0$  is the log relative rate of  $Y_t$  associated with a 1  $\mu\text{g}/\text{m}^3$  increase in the same day level of  $\text{PM}_{10}$ ,  $ns(\text{time}, df = 5/\text{year})$  is the natural cubic spline function of calendar time with 5 degrees of freedom per year,  $ns(\text{temp}, df = 7)$  is the natural cubic spline function of temperature with 7 degrees of freedom,  $I_{DOW}$  are the six indicator variables for days of the week, and  $\log(\mu_0)$  and  $\eta$  are unknown parameters.

The  $\widetilde{PM}_{10,lag0}$  values in Equation (13) were based on the following scheme. Since the degree of concavity found in the ACAPS data was 0.613, we introduced this same degree of concavity into the simulated data. The degree of concavity in the

ACAPS data was estimated by the correlation between the series of daily observed  $PM_{10}$  ( $PM_{10\_lag0}$ ) and the corresponding fitted values ( $\hat{PM}_{10\_lag0}$ ) from the additive model given by  $PM_{10\_lag0} = ns(time, df = 5/year) + ns(temp, df = 7)$ . For the simulation, a new  $PM_{10}$  series ( $PM_{10\_lag0}$ ) was generated by  $PM_{10\_lag0} = \hat{PM}_{10\_lag0} + N(0, \sigma^2)$ , where  $\sigma^2$  was chosen so that the correlation between  $PM_{10\_lag0}$  and  $\hat{PM}_{10\_lag0}$  was calculated as 0.613.

The long-term trend and seasonality data for 6-year time series was generated using

$$Trend = 1 + 0.6 \cos(2\pi \frac{day}{365.25}) + 0.4 \cos(2\pi \frac{day}{365.25}) \quad (15)$$

(1358 < day < 1732)

The factor used to rescale the trend effect in Equation (13) is set to be 0.25 (He et al., 2006). A comparison of the observed and simulated long-term and seasonal trend pattern in Figure 3 indicates the similarity of the patterns and the coherence of the peaks.

The daily mean temperature series, *temp.s*, was estimated from (14) by  $temp.s = X_n \cdot beta.temp$ , where  $X_n$  is a basis matrix generated from  $ns(temp, df = 7)$  in S-plus and *beta.temp* is a vector of the estimated coefficients for temperature in Equation (14). The comparison of the observed and simulated temperature pattern in Figure 3 indicates similarity of the patterns.

We generated 1000 sets of 2190 observations for the hospital admissions, conducted BMA analyses under AIC prior, BIC prior, and local EB estimates and calculated summary statistics. In our ACAPS the relative risk for a  $20 \mu g/m^3$  increase in the same day level of  $PM_{10}$  was estimated at 1.0003. We assumed this value to represent the true risk and compared it with the value obtained under the BMA analysis. We also investigated whether the BMA approach could correctly identify the same day  $PM_{10}$  effect when the true effect of air pollutant existed only for the same day level of  $PM_{10}$  but the model incorrectly included several  $PM_{10}$  lag

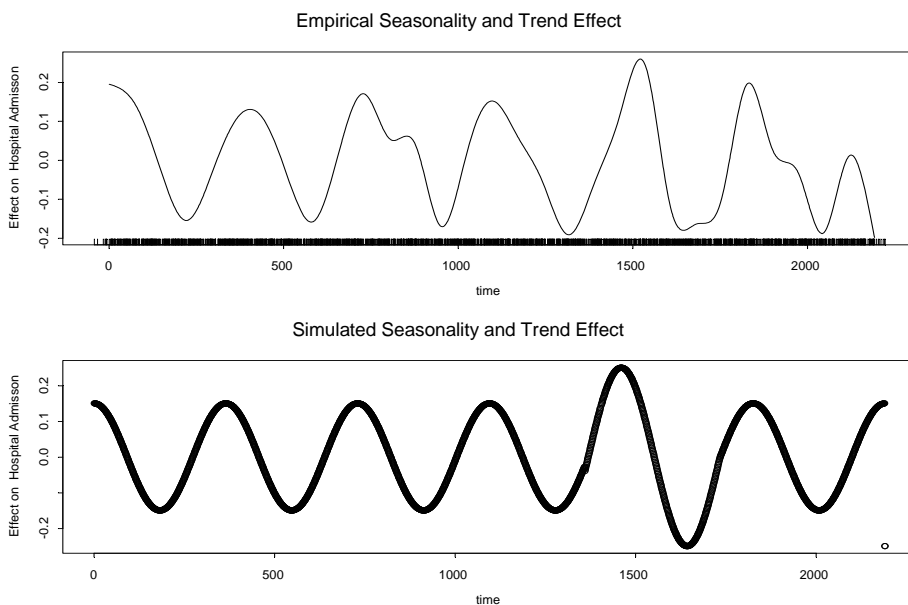
variables. Therefore, we used two models: one included all the time and temperature predictor variables in Equation (14), and the same day  $PM_{10}$  term, and the other model added  $PM_{10}$  lag variables for the five previous days together with the time and temperature variables.

With the model that included only the same day level of  $PM_{10}$  the BMA method consistently selected the same day level of  $PM_{10}$ . The estimate of relative risk was found to be close to the true value of relative risk under all three priors (Table 2.). With the model that contained same day level of  $PM_{10}$  and  $PM_{10}$  lagged by five days was used, where the underlying true model is the same day model, the BMA approach correctly selected the models that have only the same day level of  $PM_{10}$  582 to 597 times out of 1000. This showed that as  $PM_{10}$  lag variables are included the BMA approach could still have high probability to identify the true effect. However, the estimates had changed to be smaller than 1. This, we strongly believe, is due to the concavity in the data. The  $PM_{10}$  values over different days are found to be correlated and inclusion of collinear variables in regression models usually results in biased estimates.

**Table 2.** Posterior means of relative risk associated with a  $20 \mu g/m^3$  increase in the same day  $PM_{10}$  variable with 95% posterior probability intervals under BMA <sup>a</sup>

<b>PM<sub>10</sub> covariates in the fitted model</b>	<b>AIC</b>	<b>BIC</b>	<b>Local EB</b>
Same day of PM <sub>10</sub>	1.0006 (0.9975, 1.0030)	1.0009 (0.9972, 1.0036)	1.0008 (0.9973, 1.0034)
Same day and five previous days of PM <sub>10</sub>	0.9993 (0.8695, 1.1495)	0.9984 (0.8468, 1.1792)	0.9992 (0.8558, 1.1677)

<sup>a</sup> The true model used for simulation study is the same day model where the same day level of  $PM_{10}$ ,  $ns(time=5/year)$ ,  $ns(temp, df=7)$ , and day of the week variables are included in the model. The true relative risk under the assumed same day model is 1.0003.



**Figure 3.** Empirical and simulated effects of seasonal and long-term trend on hospital admissions.

## 5. Discussion

In this study, we had conducted the sensitivity analysis for BMA under AIC prior, BIC prior, and the local EB estimates in a time series study of air pollution using both the ACAPS data set and simulated data sets. An important limitation of conventional methods for analyzing air pollution time series is the failure to account for model uncertainties. Model uncertainties include several components, such as uncertainties about the variable selection procedure, uncertainties about functional forms of predictor variables, and uncertainties about the model itself. In this paper, we have considered two sources of uncertainties: (1) the uncertainties associated with the model selection procedure, which we investigated through the modeling of the ACAPS data set and (2) the uncertainty about the lagged effects. We investigated this through simulations.

We found the posterior means of the relative risk estimated by BMA under AIC prior, BIC prior, and local EB estimate were similar, ranging between 0.9980 and 1.0022 for a  $20 \mu\text{g}/\text{m}^3$  increase in all  $\text{PM}_{10}$ . Arena et al. (2006) reported a higher risk of 1.012256 for the current day level of  $\text{PM}_{10}$ , and the BMA method provides smaller estimates. The BMA estimates account more for uncertainties. We also found that the choice of prior may not be critical, at least with data similar to the ACAPS data.

Regarding the uncertainties associated with the selection of predictor variables, we found that the choice of the degree of lagging for the air pollution term was important. Results from our simulations showed that if the lag variables of  $\text{PM}_{10}$  were incorrectly considered in the model, the estimates of relative risk could change and in our case decreased. We attribute this to the concavity problem that results from the inclusion of lag variables of  $\text{PM}_{10}$ . Because these lagged predictor variables are collinear, GAMs, which are based on the backfitting algorithm, can present instability with respect to the order of variables or to the subset of variables in the fitting process. Future research may apply other methods, such as projection methods, which perform a nonlinear transformation from the space of the inputs and then a linear transformation from this new space, that are not affected by the collinearity into BMA. It should also be noted that these results are based on the simulation of a particular data set and degree of concavity; other data sets may show changes of a greater or lesser degree and in either direction (a risk that is biased upwards or downwards). These results indicate the importance of selecting the correct degree of lagging for variables, not based on only maximizing the likelihood, but by considering the amount of concavity, and biological plausibility. It is also important to investigate whether BMA could correctly identify a true, multiday lagged effect which is beyond the scope of the present paper. We leave this for future work. In these analyses we have not considered the uncertainties associated with the functional form of the model. This source of uncertainty is as important as the others, and possibly the most difficult to assess.

Regarding the interpretation of the pollutant effect arising from a BMA analysis when BMA is considered to be more suited for prediction rather than interpretation of a specific regression coefficient (Thomas et al., 2007), we note that that the assumed invariance of the interpretation of this effect in each competing BMA model poses no problem as we have only one pollutant in our model.

## Acknowledgements

This research was partially supported by ExxonMobil Agreement # A173647 to the University of Pittsburgh.

## References

- Arena, V.C., Mazumdar, S., Zborowski, J.V., Talbott, E.O., He, S., Chuang, Y.H., Schwerha, J.J., 2006. A retrospective investigation of  $\text{PM}_{10}$  in ambient air and cardiopulmonary hospital admissions in Allegheny County, Pennsylvania: 1995-2000. *Journal of Occupational and Environmental Medicine* 48, 38-47.
- Clyde, M., 2000. Model uncertainty and health effect studies for particulate matter. *Environmetrics* 11, 745-763.
- DiMatteo, I., Genovese, C.R., Kass, R.E., 2001. Bayesian curve-fitting with free-knot splines. *Biometrika* 88, 1055-1071.
- George, E.I., Foster, D.P., 2000. Calibration and empirical Bayes variable selection. *Biometrika* 87, 731-747.
- Hansen, M.H., Yu, B., 2003. Minimum description length model selection criteria for generalized linear models: Science and Statistics: Festschrift for Terry Speed. 145-163. Fountain Hills, AZ, IMS Press. IMS Lecture Notes - Monograph Series.
- Hansen, M. H., Yu, B., 2001. Model selection and the principle of minimum description length. *Journal of the American Statistical Association* 96, 746-774.
- He, S., Mazumdar, S., Arena, V.C., 2006. A comparative study of the use of GAM and GLM in air pollution research. *Environmetrics* 17, 81-93.
- Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: A tutorial. *Statistical Science* 14, 382-401.
- Holmes, C.C., Mallick, B.K., 2003. Generalized nonlinear modeling with multivariate free-knot regression splines. *Journal of the American Statistical Association* 98, 352-368.
- Lee, D., Shaddick, G., 2008. Modelling the effects of air pollution on health using Bayesian dynamic generalised linear models. *Environmetrics* 19, 785-804.
- Liu, Y., Guo, H., Mao, G.Z., Yang, P.J., 2008. A Bayesian hierarchical model for urban air quality prediction under uncertainty. *Atmospheric Environment* 42, 8464-8469.
- Madigan, D., Raftery, A.E., 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 89, 1535-1546.
- Nikolov, M.C., Coull, B.A., Catalano, P.J., Godleski, J.J., 2007. An informative Bayesian structural equation model to assess source-specific health effects of air pollution. *Biostatistics* 8, 609-624.
- Ramsay, T.O., Burnett, R.T., Krewski, D., 2003. The effect of concavity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology* 14, 18-23.
- Schwartz, J., 1993. Air pollution and daily mortality in Birmingham, Alabama. *American Journal of Epidemiology* 137, 1136-1147.
- Smith, R.L., Davis, J.M., Sacks, J., Speckman, P., Styer, P., 2000. Regression models for air pollution and daily mortality: Analysis of data from Birmingham, Alabama. *Environmetrics* 11, 719-743.
- Thomas, D.C., Jerrett, M., Kuenzli, N., Louis, T.A., Dominici, F., Zeger, S., Schwarz, J., Burnett, R.T., Krewski, D., Bates, D., 2007. Bayesian model averaging in time-series studies of air pollution and mortality. *Journal of Toxicology and Environmental Health-Part A-Current Issues* 70 311-315.
- Wordley, J., Walters, S., Ayres, J.G., 1997. Short term variations in hospital admissions and mortality and particulate air pollution. *Occupational and Environmental Medicine* 54, 108-116.