

**A COMPARATIVE STUDY OF INFERENTIAL
PROCEDURES FOR AIR POLLUTION HEALTH
EFFECTS RESEARCH**

by

Ya-Hsiu Chuang

BBA Statistics, Tunghai University, Taiwan, 1999

MPS Applied Statistics, Cornell University, 2002

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH
GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Ya-Hsiu Chuang

It was defended on

July 28th, 2009

and approved by

Dissertation Advisor:

**Sati Mazumdar, PhD, Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh**

Taeyoung Park, PhD, Assistant Professor, Department of Statistics
Graduate School of Arts and Sciences, University of Pittsburgh

Gong Tang, PhD, Assistant Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Vincent C. Arena, PhD, Associate Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Howard E. Rockette, PhD, Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Copyright © by Ya-Hsiu Chuang

2009

A COMPARATIVE ANALYSIS OF INFERENTIAL PROCEDURES FOR AIR POLLUTION HEALTH EFFECT STUDIES

Ya-Hsiu Chuang, PhD

University of Pittsburgh, 2009

Generalized additive model (GAM) with natural cubic splines (NS) has been commonly used as a standard analytical tool in time series studies of health effects of air pollution. Standard model selection procedures used in GAM ignore the uncertainty in model fitting. This may lead to biased estimates of the health effects, in particular lagged effects. Moreover, the degrees of smoothing to adjust for time-varying confounders estimated from data-driven methods were found to give biased estimates. We applied Bayesian model averaging (BMA) approach to account for model uncertainty and proposed also a generalized linear mixed model with natural cubic splines (GLMM + NS) to adjust for time-varying confounders. As the posterior model probability derived from BMA contains a hyperparameter to account for model uncertainty and has potential usefulness in this type of studies, we first conducted a sensitivity analysis with simulation studies for BMA with different calibrated hyperparameters. Our results indicated the importance of selecting the optimum degree of lagging for variables, not based on only maximizing the likelihood, but by considering the possible effects of lagging and biological plausibility. Our proposed model, GLMM + NS, was found to produce more precise estimates of the health effects of current day level of PM_{10} than the commonly used generalized linear models with natural cubic splines (GLM + NS) in our simulation studies. However, more in depth analyses with special attention to inferential procedures in readily available software are needed to have any definitive conclusion about the performance of our proposed model. An

illustrative example is provided using data from the Allegheny County Air Pollution Study (ACAPS) where the quantity of interest was the relative risk of cardiopulmonary hospital admissions for a $20 \mu\text{g}/\text{m}^3$ increase in PM_{10} for the current day and five previous days. Assessing the effect of air pollution on human health is an important public health problem. There are some inconsistencies in the literature as to the magnitude of this effect. The proposed statistical methods are expected to better characterize the true effect of air pollution.

TABLE OF CONTENTS

PREFACE.....	XI
1.0 INTRODUCTION.....	1
1.1 CURRENT MODELING ISSUES IN TIME-SERIES STUDIES OF AIR POLLUTION.....	2
1.2 OVERVIEW AND AIMS.....	3
1.2.1 Aim 1: Bayesian Model Averaging Approach in Health Effects Studies: Sensitivity Analyses Using PM₁₀ and Cardiopulmonary Hospital Admissions in Allegheny County, Pennsylvania and Simulated Data.....	4
1.2.2 Aim 2: Generalized Linear Mixed Models Approach in Time Series Studies of Air Pollution.....	6
2.0 BAYESIAN MODEL AVERAGING APPROACH IN HEALTH EFFECTS STUDIES: SENSITIVITY ANALYSES USING PM₁₀ AND CARDIOPULMONARY HOSPITAL ADMISSIONS IN ALLEGHENY COUNTY, PENNSYLVANIA AND SIMULATED DATA	7
2.1 ABSTRACT.....	8
2.2 INTRODUCTION	9
2.3 BAYESIAN MODEL AVERAGING.....	11
2.3.1 Bayesian Model Averaging (BMA)	11

2.3.2	Implementation of BMA	12
2.4	APPLICATION BMA METHOD TO THE ACAPS DATA	16
2.4.1	ACAPS Data and Model Fitted.....	16
2.4.2	Bayesian Model Averaging Analysis.....	18
2.5	SIMULATION STUDY.....	21
2.6	DISCUSSION.....	26
2.7	ACKNOWLEDGEMENTS	28
3.0	GENERALIZED LINEAR MIXED MODELS APPROACH IN TIME SERIES STUDIES OF AIR POLLUTION.....	29
3.1	ABSTRACT.....	30
3.2	INTRODUCTION	31
3.3	REVIEW OF GENERALIZED ADDITIVE MODELS AND CHOICES OF SMOOTH FUNCTIONS	33
3.3.1	Generalized Additive Model (GAM)	33
3.3.2	Choices of Smooth Functions	34
3.3.3	Estimation of GAMs.....	36
3.4	GENERALIZED LINEAR MIXED MODELS WITH NATURAL CUBIC SPLINES	39
3.5	SIMULATION STUDY.....	43
3.6	APPLICATION OF GENERALIZED LINEAR MIXED MODELS WITH NATURAL CUBIC SPLINES TO THE ACAPS DATA.....	46
3.6.1	ACAPS Data	46
3.6.2	Models Fitted	47

3.7	CONCLUSION AND DISCUSSION	51
4.0	CONCLUSION AND DISCUSSION.....	52
	BIBLIOGRAPHY.....	54

LIST OF TABLES

Table 1 Summary of the posterior distribution of relative risk associated with a $20 \text{ ug}/\text{m}^3$ increase in all PM_{10} under BMA using ACAPS data set	21
Table 2 Posterior means of relative risk associated with a $20 \text{ ug}/\text{m}^3$ increase in all PM_{10} and their 95% posterior probability intervals under BMA using simulated data set [†]	26
Table 3 Empirical bias of $\hat{\beta}_{\text{PM}_{10_lag0}}$ with empirical standard deviations in the parentheses	45
Table 4 Summary for the fixed effect estimate of PM_{10_lag0}	50

LIST OF FIGURES

Figure 1 Plots of model space.....	19
Figure 2 Distribution of relative risks using BMA approach given PM_{10} is included	20
Figure 3 Empirical and simulated effect of seasonal and long-term trend on hospital admissions	24
Figure 4 Empirical and simulated effect of temperature on hospital admissions	24

PREFACE

I would like to express my gratitude for the opportunities and training that have been provided to me during my doctoral studies. The training from the course work and being a graduate student researcher (GSR) have allowed me to apply what I have learned in class to the real world. This research began with the Exxon project with Exxon Agreement # A173647 and has continued as my research topic.

With the deepest gratitude, I wish to thank my advisor, Dr. Sati Mazumdar, for her magnificent support and patience. Dr. Mazumdar has been my mentor and GSR supervisor since 2007. Under her surveillance, I deeply appreciate her great support, guidance and tutelage.

I would like to acknowledge and express my gratitude to my committee members: Dr. Taeyoung Park, Dr. Gong Tang, Dr. Vincent Arena, and Dr. Howard Rockette for providing me with their valuable feedback, comments and advice. I benefited a lot from them.

I would also like to express my gratitude to Dr. Mark Nicolich, Statistician in Exxon Mobil Biomedical Sciences, Inc., for being my external advisor and for giving me comments and advice.

In the end, I would like to thank my family for giving me all of their support and keeping their faith in me. They are the most important people in my life.

1.0 INTRODUCTION

The early air pollution studies in Meuse Valley in Belgium in 1930, Donora, Pennsylvania in 1948, and London in 1952 showed a large impact of air pollution on public health. This drew the attention and motivated the respective governments to initiate research in this area and to enact legislation to improve the air quality. The United States Congress enacted the Clean Air Act (CAA) in 1970 and the regulations were developed by the Environmental Protection Agency (EPA) to review the National Ambient Air Quality Standards (NAAQS) for pollutants. The EPA compiles and assesses the research related to the impact of pollutants on public health. The EPA also evaluates the policy implications and makes recommendations for policy options. Based on the comments from public interest groups, private industry and recommendations from the EPA, the Clean Air Scientific Advisory Committee (CASAC), a congressionally mandated panel of scientific experts, makes final recommendations and then the EPA proposes changes to regulatory standards. To date, several epidemiological study designs have been used to investigate the health effects of air pollution on public health. Most of the studies are based on either cohort or time-series approaches. The cohort studies that follow a fixed group of individuals over a long time span are used to compare long-term average pollution levels and adjust health outcomes, mostly across geographic locations. The individual level information such as smoking status, race, and body mass index can be accounted for in the cohort studies. Time-series studies are used to assess the effects of short-term changes in air pollution on acute

health effects by estimating the associations between daily variations in air pollution and counts in health outcomes.

1.1 CURRENT MODELING ISSUES IN TIME-SERIES STUDIES OF AIR POLLUTION

Generalized additive model (GAM) (Hastie and Tibshirani, 1990) has been commonly used as a standard analytical tool in time-series studies of air pollution. GAM allows flexibility in the specifications of the non-linear functions of variables in the model, where the non-linear relationships exist in the long-term trends, seasonality and temperature with health outcomes. Like common model building procedures, GAM follows the standard rule that selects a subset of predictor variables according to their statistical significance levels. As the effects of ambient air levels of a pollutant could last for more than one day, determination of the lagged effects of air levels of a pollutant on cardiopulmonary distress becomes important. However, standard model selection procedures determine optimum degree of lagging on the effects of air pollution by their statistical significance levels. This leads to the problem of not accounting for model uncertainties associated with variable selection procedures.

Another issue with GAM modeling is the degrees of smoothing that are used for the smooth functions to adjust for the long-term trends and seasonality and temperature effects. Data-driven methods used to determine the degrees of smoothing of the smooth function are found to give less accurate estimates on the health effects of air pollution as concurvity increases, where concurvity is a non-parametric analogue of multicollinearity (Peng *et al.* 2006). Furthermore, the derived degrees of smoothing are assumed to be fixed over the study period.

Assumptions of the fixed degrees of smoothing may fail to capture the true relationship of the non-linear trend, especially in the scenario that the curve is wiggly in some areas and smooth in others.

While research regarding the uncertainty has been conducted (Clyde 2000), it did not consider the lag lengths of the effects of air pollution. In addition, advanced modeling approaches in dealing with the smoothing problem on the estimation of the health effects of air pollution are still limited. This prompts our interest in the investigation of estimation of the health effects of air pollution with new inferential procedures such as Bayesian model averaging and generalized linear mixed models. This dissertation includes theories and simulation studies accompanied by illustrative examples from one air pollution study to elucidate the proposed approaches.

1.2 OVERVIEW AND AIMS

The objective of our research was to examine the estimation of the health effects of air pollution with respect to the parameter estimates and their standard errors. Two aims are specified as follows:

Aim 1: Conduct sensitivity analyses using Bayesian model averaging approach to account for the uncertainties resulting from the variable selection procedure in determining the number of lagged effects of air pollution, accompanied by the simulation study.

Aim 2: Develop a generalized linear mixed model with natural cubic splines to handle the problem of fixed degrees of smoothing.

This dissertation is organized into two self-contained manuscripts presented in Chapters 2 and 3. Each manuscript addresses one specific aim. Conclusions and discussions are presented in Chapter 4.

1.2.1 Aim 1: Bayesian Model Averaging Approach in Health Effects Studies: Sensitivity Analyses Using PM₁₀ and Cardiopulmonary Hospital Admissions in Allegheny County, Pennsylvania and Simulated Data

One of the major statistical issues in the study of air pollution is the lagged effects of air pollutant (e.g. PM₁₀: particulate matter with a mean aerodynamic diameter of 10 μm or less) on cardiopulmonary distress. As the effects of PM₁₀ on cardiopulmonary distress can last for more than one day, how long the health effects of air pollution usually last has drawn attention. Smith *et al.* (2000) applied standard model selection procedures to determine the optimum degree of lagging of PM₁₀ and found that none of the lag variables was statistically significantly associated with non-accidental elderly mortality using Birmingham, AL, data from 1985 through 1988. Wordley *et al.* (1997) used Birmingham, UK, data from 1992 to 1994 and included PM₁₀ 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 effects of PM₁₀ in the model. Statistically significant associations of these variables with all respiratory hospital admissions were found. Additionally, Schwartz (1993) used the average of PM₁₀ for the three previous days and found a statistically significant effect between PM₁₀ and non-accidental elderly mortality. All these studies showed that the statistical significance of the health effects of air pollution varies with different formulations on the PM₁₀ variables. This variation may be due to ignoring uncertainties associated with variable selection procedures in standard model selection approaches.

Bayesian model averaging (BMA) has the advantage of accounting for the model uncertainties existing in the variable selection procedures. Clyde (2000) applied BMA to study the health effects of air pollution and developed a class of objective prior distributions to provide weights for the models to effectively account for the uncertainties. The posterior model probabilities derived from BMA can be calibrated to different classical model selection criteria, such as Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC), through the choice of hyperparameter g . As these two criteria use information on the number of observations in the study and the number of parameters in the models, they ignore other information contained in the data. Therefore, we adapted the local Empirical Bayes (EB) approach to estimate the hyperparameter from the data, where different models have different estimates of g . A sensitivity analysis for BMA with different calibrated hyperparameters was conducted to compare the estimates of the relative risks of cardiopulmonary hospital admissions. We performed a simulation study to investigate whether BMA could correctly select the true models that include the lag effects of air pollutant.

Manuscript/Presentation Status: This work is the winner of the 2009 Student Paper Competition on the Risk Analysis Section of the American Statistical Association (ASA) and will be presented in Joint Statistical Meetings (JSM) 2009. A manuscript entitled “Bayesian Model Averaging Approach in Health Effects Studies: Sensitivity Analyses Using PM_{10} and Cardiopulmonary Hospital Admissions in Allegheny County, Pennsylvania and Simulated Data” has been submitted to *Environmetrics*. Chapter 2 replicates this manuscript.

1.2.2 Aim 2: Generalized Linear Mixed Models Approach in Time Series Studies of Air Pollution

The model fitting of GAM requires the specification of the degrees of smoothing for the smooth functions. The degrees of smoothing were found to have more effects than the choices of smoothers on the estimation of the effects of air pollution (Rupert *et al.* 2003, Peng *et al.* 2006). Peng *et al.* (2006) showed that the degrees of smoothing determined by data-driven methods that optimize the predictivity of the data series do not give accurate estimates on the effects of air pollution, especially under high concavity, which occurs in much of the air pollution data (Peng *et al.* 2006). While it was suggested that giving larger degrees of smoothing than what are estimated from the data could lead to less biased estimates of the effects of air pollution, it can be questionable whether the larger degrees of smoothing could correctly capture the true curve. Furthermore, the degrees of smoothing derived from these methods are assumed to be fixed over the study period. This assumption can be violated if the true underlying smooth function is wiggly in some subsets of the study period and smoothly in others.

Instead of adapting fixed degrees of smoothing, we proposed a generalized linear mixed model with natural cubic splines (GLMM + NS) to allow the degrees of smoothing to vary locally to capture the shape of the true effects. We compared the proposed methods to the existing methods through an illustrative example using data from the Pittsburgh area (Arena *et al.*, 2006) and a simulation study to examine the performance of the proposed method with respect to the parameter estimates and its standard errors.

Manuscript/Presentation Status: This research has not been presented. The manuscript in progress is presented in Chapter 3.

**2.0 BAYESIAN MODEL AVERAGING APPROACH IN HEALTH EFFECTS
STUDIES: SENSITIVITY ANALYSES USING PM₁₀ AND CARDIOPULMONARY
HOSPITAL ADMISSIONS IN ALLEGHENY COUNTY, PENNSYLVANIA AND
SIMULATED DATA**

Ya-Hsiu Chuang¹, Mark J. Nicolich² and Sati Mazumdar¹

¹Department of Biostatistics, School of Public Health, University of Pittsburgh, Pittsburgh,

Pennsylvania, USA

² Lambertville, New Jersey, USA

2.1 ABSTRACT

Determining the lagged effects of ambient air levels of a pollutant on cardiac distress is important in health effects studies. Standard model selection procedures where a set of predictor variables is selected ignore the associated uncertainties and may lead to overestimation of effects. Bayesian model averaging approach accounts for model uncertainty by combining information from all possible models. Zellner's g -prior containing a hyperparameter g can account for model uncertainty and has potential usefulness in this endeavor. We conducted a sensitivity analysis for Bayesian model averaging with different calibrated hyperparameter g , viz., Akaike Information Criterion (AIC) prior, Bayes Information Criterion (BIC) prior, and Local Empirical Bayes estimates. Data from the Allegheny County Air Pollution Study (ACAPS) and simulated data sets were used. Our main quantity of interest was the relative risk of cardiopulmonary hospital admissions for a $20 \mu\text{g}/\text{m}^3$ increase in PM_{10} values for the current day and five previous days. Results showed that the posterior means of the relative risk and 95% posterior probability intervals were close to each other under different prior choices: 0.9936 (0.9861, 1.0085) with AIC prior, 0.9913 (0.9033, 1.0987) with BIC prior, and 0.9926 (0.9111, 1.0905) with local Empirical Bayes estimate. Simulation results were consistent with these findings.

Keywords: lagged effects; Bayesian model averaging; hyperparameters; AIC; BIC; local Empirical Bayes

2.2 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_{10} : particulate matter with a mean aerodynamic diameter of $10\mu m$ or less) on cardiopulmonary distress. Theoretically, the effect of PM_{10} on cardiopulmonary distress can last for more than one day. Therefore, it is important to find exactly how long this effect usually lasts. Birmingham, AL, data from 1985 through 1988 have been used to study this effect in Smith *et al.* (2000) and Schwartz (1993). Smith *et al.* (2000) applied standard model selection procedures to determine the number of lag variables of different lengths for PM_{10} and found that none of the lag variables were statistically significantly associated with non-accidental elderly mortality. Schwartz (1993) used the average of PM_{10} for the three previous days and found a

statistically significant effect between PM_{10} and non-accidental elderly mortality. In the analysis of Allegheny County Air Pollution Study (ACAPS) where data for Pittsburgh, PA, from 1995 to 2000 were used, only the same day level of PM_{10} was found to have an effect on the cardiopulmonary hospital admissions (Arena *et al.* 2006). Wordley *et al.* (1997) used Birmingham, UK, data from 1992 to 1994 and included PM_{10} 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_{10} in the model. Statistically significant associations of these variables with all respiratory hospital admissions were found. Therefore, the problem with standard model selection approaches lies in not taking into account uncertainties associated with variable selection procedures.

Bayesian model averaging (BMA), developed by Kass and Raftery (1995), provides an approach to take into account model uncertainty by combining information from all possible models and obtaining a weighted average for the quantity of interest over these models. 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 suggested providing estimates for several prior distributions, i.e., from several choices of the hyperparameter, thus providing a sensitivity analysis (Clyde 2000). When the results from the sensitivity analysis are not consistent, further studies need to be performed.

Section 2.3 presents a brief description of BMA and methods for its implementation. An illustrative example using ACAPS data is presented in Section 2.4. ACAPS is a time-series study to investigate the effect of ambient air level of PM_{10} on daily cardiopulmonary hospital admissions in elderly residents of Allegheny County from 1995 through 2000 (Arena *et al.* 2006). A simulation study is given in Section 2.5 followed by a discussion in Section 2.6.

2.3 BAYESIAN MODEL AVERAGING

2.3.1 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 given the data. Let Λ 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 Λ given the data Y can be written as

$$\Pr(\Lambda | Y) = \sum_{m=1}^K \Pr(\Lambda | Y, M_m) \Pr(M_m | Y) , \quad (2.1)$$

where M_m is the m^{th} model under consideration 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 (2.1) is the predictive distribution of Λ given a particular model M_m and the data, and the second term is the posterior probability of model M_m given the data.

2.3.2 Implementation of BMA

The predictive distribution of Λ given a particular model M_m and Y in (2.1) is given by

$$\Pr(\Lambda | Y, M_m) = \int \Pr(\Lambda | \beta_m, M_m, Y) \Pr(\beta_m | M_m, Y) d\beta_m, \quad (2.2)$$

where β_m is the vector of regression coefficients for the model M_m . For regression models where the integration can be of high dimensional, (2.2) may not provide any closed form solutions. In such situations, the maximum likelihood estimate (MLE) of β_m can be used giving

$$\Pr(\Lambda | Y, M_m) \approx \Pr(\Lambda | \hat{\beta}_m, Y, M_m). \quad (2.3)$$

This approach was found to give an excellent approximation in time series regression problem (Taplin 1993, Taplin and Raftery 1994).

The posterior probability for model M_m is given by

$$\Pr(M_m | Y) = \frac{\Pr(Y | M_m) \Pr(M_m)}{\sum_{j=1}^K \Pr(Y | M_j) \Pr(M_j)} \quad (2.4)$$

where,

$$\Pr(Y | M_m) = \int \Pr(Y | \beta_m, M_m) \Pr(\beta_m | M_m) d\beta_m, \quad (2.5)$$

$\Pr(\beta_m | M_m)$ is the prior density of β_m under model M_m , and $\Pr(M_m)$ is the prior density. In order to derive the posterior model probability, these prior densities need to be specified in advance. We take $\Pr(M_m)$ as the uniform distribution. One of the most common prior choices for β_m is Zellner's g -prior (Zellner, 1986) which is defined as

$$\beta_m | M_m, \phi \sim N\left(\beta, \frac{g}{\phi} (X_m^T X_m)^{-1}\right), \quad (2.6)$$

where g is the hyperparameter and ϕ is the precision parameter with

$$\Pr(\phi | M_m) \propto \frac{1}{\phi}. \quad (2.7)$$

By adapting (2.6) and (2.7) in (2.5), we get

$$\Pr(Y | M_m) = \int \int \Pr(Y | \beta_m, M_m) \Pr(\beta_m | M_m, \phi) \Pr(\phi | M_m) d\phi d\beta_m. \quad (2.8)$$

As (2.8) can be of high dimensional and may not have an analytic form, one can use Laplace methods to approximate it (Tierney and Kadane, 1986) by

$$\int \exp(f(u)) du \approx (2\pi)^{p/2} |A|^{1/2} \exp(f(u^*)),$$

where u^* is the value of u at which f attains its maximum, and A is the negative inverse Hessian of f evaluated at u^* . Hence,

$$\Pr(Y | M_m) \approx (2\pi)^{\frac{d_m}{2}} |\Psi_k|^{-\frac{1}{2}} \Pr(Y | \tilde{\beta}_m, M_m) \Pr(\tilde{\beta}_m | M_m) \quad (2.9)$$

where, d_m is the dimension of β_m , $\tilde{\beta}_m$ is the posterior mode of β_m , and Ψ_k is the negative inverse Hessian of $\log\{\Pr(Y | \beta_m, M_m) \Pr(\beta_m | M_m)\}$, evaluated at $\beta_m = \tilde{\beta}_m$ (Raftery, 1996).

Using (2.9) in (2.4), the posterior model probability is given by

$$\Pr(M_m | Y) = \frac{\exp(0.5 * (D_m - d_m \log(g)))}{\sum_{j=1}^K \exp(0.5 * (D_j - d_j \log(g)))}, \quad (2.10)$$

where D_m is the model deviance, d_m is the dimension of β_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 (2.1) 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. When the number of predictor variables is more than 30, Hoeting *et al.* (1999) suggested using the leaps and bounds method to eliminate variables, where the leaps and bounds algorithm is an algorithm that provides top models based on the residual sum of squares (Furnival and Wilson, 1974). The posterior mean and variance of Λ are given by (Kass and Raftery, 1995)

$$E(\Lambda | Y) = \sum_{m=1}^K E(\Lambda | Y, M_m) \Pr(M_m | Y) , \quad (2.11)$$

$$Var(\Lambda | Y) = \sum_{m=1}^K (Var(\Lambda | Y, M_m) + (E(\Lambda | Y, M_m))^2 \cdot \Pr(M_m | Y) - E(\Lambda | Y)^2) . \quad (2.12)$$

In the previous formulation, Zellner's g -prior was used as the prior distribution of the parameters β_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 through the choice of hyperparameter g (Clyde 2000). In addition, an Empirical Bayes (EB) approach was developed to provide adaptive estimates of g . The local EB approach (Hansen and Yu, 2001, 2003) 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

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

MLE of β_m and d_m is the dimension of model M_m (Hansen and Yu, 2003).

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 PM_{10} variables ($PM_{10_lag0}, \dots, PM_{10_lagq}$), the relative risks for each model were given by

$$\Lambda = \exp[20 * (\beta_{PM_{10_lag0}} + \beta_{PM_{10_lag1}} + \dots + \beta_{PM_{10_lagq}})] \quad (2.13)$$

where q is the lag length of the PM_{10} .

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

$$\log(\Lambda) | M_m, Y \sim N(20 * (\beta_{PM_{10_lag0}} + \dots + \beta_{PM_{10_lagq}}), \sigma_{RR}^2), \quad (2.14)$$

where $\sigma_{RR}^2 = 20^2 (1^T \Sigma_{\beta_{PM}|M_m} 1)$ with $1^T = (1, \dots, 1)$ of dimension q and $\Sigma_{\beta_{PM}|M_m}$ is the covariance matrix for the PM_{10} variables under model M_m derived from the Fisher information under model M_m (Clyde, 2000).

2.4 APPLICATION BMA METHOD TO THE ACAPS DATA

2.4.1 ACAPS Data and Model Fitted

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 (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 temperature data were used as the meteorological data in our study. Ambient air levels of a criteria pollutant (PM_{10}) were recorded in every hour for each of the 8 monitoring sites. The mean of the site-specific daily average PM_{10} values across all monitoring was used. Since only two sets of data out of 2192 were missing on dates 03/24/1998 and 11/04/1998, they were ignored and resulted in a total of 2190 observations for our data analysis.

The selected predictor variables in this paper included the levels of PM_{10} for same day and lagged up to five days ($PM_{10_lag0}, \dots, PM_{10_lag5}$), the daily mean temperature (MNTTP), 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 in the fit of GAMs. This leads to the

suggestion of fitting GLMs with natural cubic spline. 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 concavity exists in the data. Therefore, we have limited our analyses to models using a GLM with natural cubic splines.

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. The degrees of smoothing for the long-term trend and seasonality were chosen from the fitted model that has the smallest AIC, where the smaller AIC indicates 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 previous generalized linear model and repeated the same procedure to find the degrees of smoothing. This resulted in 5 degrees of freedom per year for long-term trend and seasonality, and 7 degrees of freedom for daily mean temperature. Therefore, the fitted GLM with natural cubic spline that is used in this paper is given by:

$$\begin{aligned}
Y_t &\sim \text{Poisson}(\mu_t) \\
\log(\mu_t) &= \alpha_0 + \beta_0 PM_{10_lag0} + \dots + \beta_5 PM_{10_lag5} \\
&\quad + ns(\text{time}, df = 5 / \text{year}) + ns(MNTP, df = 7) + \eta * I_{DOW}, \tag{2.15}
\end{aligned}$$

where Y_t is the counts of daily cardiopulmonary hospital admissions, which follows a Poisson distribution with mean μ_t , $PM_{10_lag0}, \dots, PM_{10_lag5}$ 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, $ns(MNTP, df = 7)$ is the natural cubic spline function

of temperature with 7 degrees of freedom, and I_{DOW} are the six indicator variables for days of the week.

2.4.2 Bayesian Model Averaging Analysis

The model given in (2.15) includes 49 predictor variables resulting in 2^{49} possible models. We first reduced the number of predictor variables to 30 by the leaps and bounds method, where the eliminated variables were the levels of PM_{10} on 4-day and 5-day lags, the functional terms of the smooth functions for long-term trend and seasonality with knots placed at the 1st, 2nd, 5th, 6th, 14th, 15th, 17th, 23rd, 26th, 27th, 28th, 29th, and 30th joint point, the functional terms of smooth function for temperature with knots placed at the 1st and 2nd joint point, and the indicator variables for day of the week on Sunday and Monday. Among all of the PM_{10} variables considered in (2.15), the variables of PM_{10} on the same day and lagged by up to three days were selected. Then Occam's Window was applied to find the data-supported model through the modified *bic.glm* package in R. It identified the first 150 models that have 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 AIC prior, BIC prior, and local EB estimate (Clyde 2002). Model matrices have the advantages of allowing us to visually identify which variables have more critical influence on the outcome variable in which we are interested. The x -axis for a model matrix represents the selected models ordered from the best to the worst (moving from left to right) based on the posterior model probabilities and the y -axis shows the predictor variables under consideration from leaps and bounds method. The top 25 models selected from BMA under AIC, BIC, and local EB estimate are shown in Figure 1. The names of the predictor variables starting with “ nt ” and “ $nmntp$ ” on the y -axis of a model matrix

represent the smooth functions for long-term trend and seasonality and for daily mean temperature, respectively. The number followed by “*nl*” and “*nmntp*” is the knot number specified through the degree of freedom of the natural cubic spline. The dark blocks in the matrix represent the selected predictor variables under a given model. The same day level of PM₁₀ was found in most of the data-supported models for BMA under AIC prior, BIC prior, and local EB estimate (Figure 1). Additionally, the posterior model probabilities under BIC prior and local EB estimate were shown to be much higher than those under AIC prior.

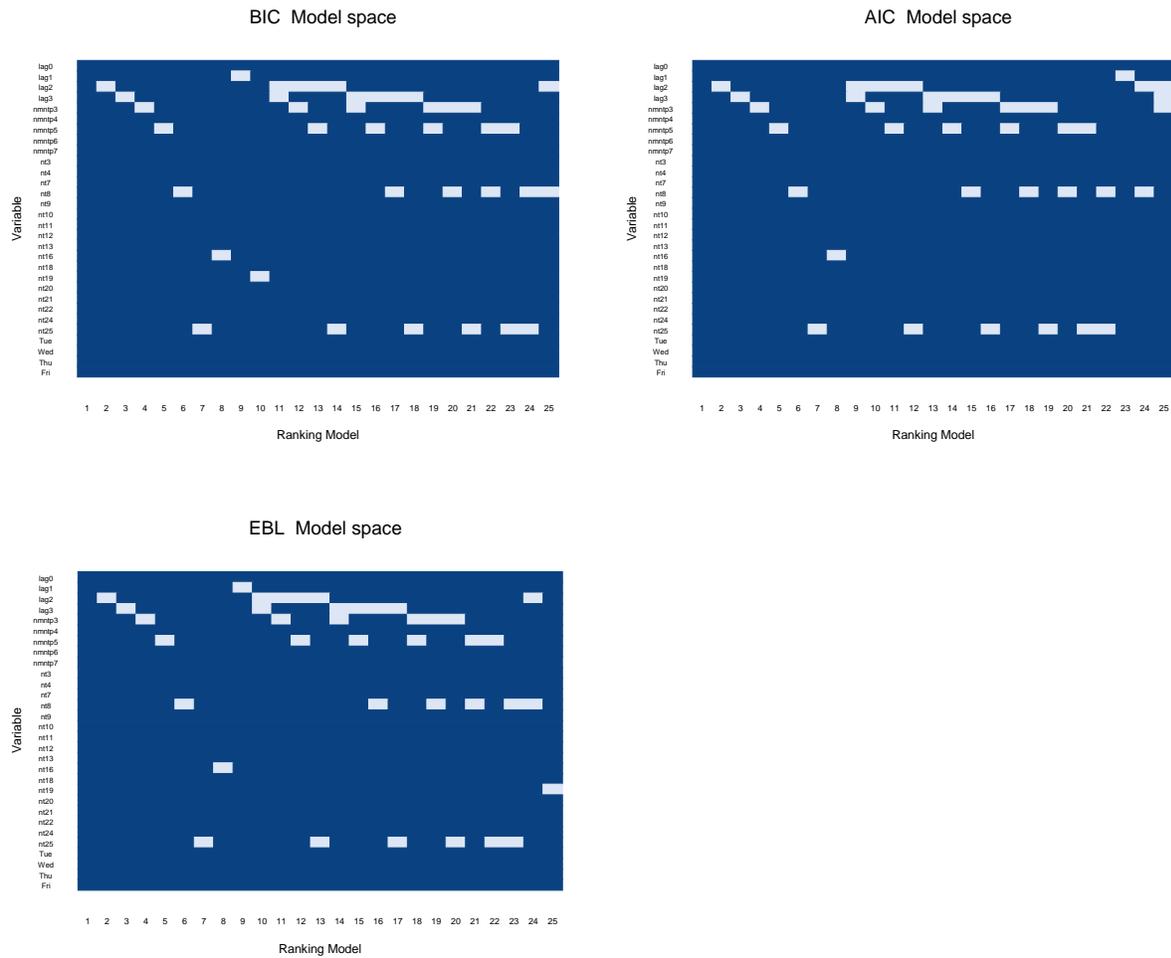


Figure 1 Plots of model space

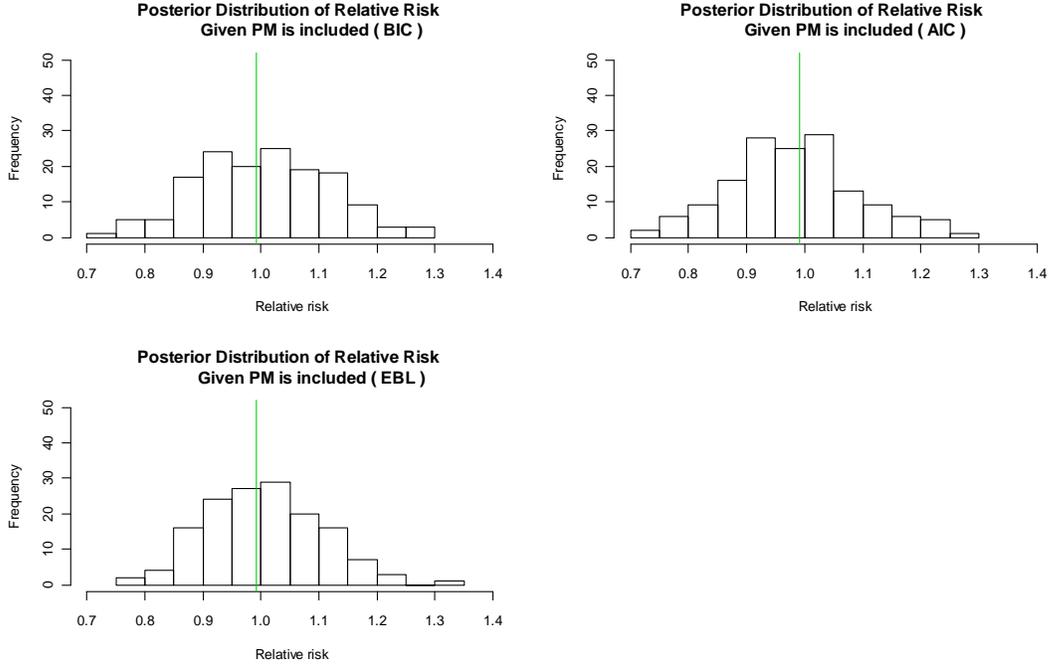


Figure 2 Distribution of relative risks using BMA approach given PM_{10} is included

Under model (2.15), according to (2.13) and (2.14), the relative risks for each model based on a $20 \mu g/m^3$ increase in all the PM_{10} variables ($PM_{10_lag0}, \dots, PM_{10_lag5}$) were given by

$$\Lambda = \exp[20 * (\beta_{PM_{10_lag0}} + \beta_{PM_{10_lag1}} + \dots + \beta_{PM_{10_lag5}})],$$

and the posterior distribution for the relative risk given M_m was

$$\log(\Lambda) | M_m, Y \sim N(20 * (\beta_{PM_{10_lag0}} + \dots + \beta_{PM_{10_lag5}}), \sigma_{RR}^2),$$

where $\sigma_{RR}^2 = 20^2 (1^T \Sigma_{\beta_{PM} | M_m} 1)$ with $1^T = (1, 1, 1, 1, 1)$.

The posterior means of the relative risk and the 95% posterior probability intervals derived from (2.11) and (2.12) were reported in Table 1. Based on a $20 \mu g/m^3$ increase in all the PM_{10} variables ($PM_{10_lag0}, \dots, PM_{10_lag5}$), the posterior means of the relative risk ranging from 0.991 to 0.994 for BMA under AIC prior, BIC prior, and local EB estimate. The consistent estimates for the posterior means of the relative risk for BMA under either of the prior choices

could be because the less important predictor variables had been screened out during the variable reduction steps of the leaps and bounds method and the predictor variables remaining for BMA were more essential. The posterior probability intervals for BMA with the BIC prior and local EB estimate were found wider than these for BMA with the AIC prior. As BIC prior and local EB estimate utilize the information from the data to estimate the hyperparameter g , this could lead to the greater levels of uncertainty and therefore results in the wider posterior probability intervals. The posterior distributions of the relative risk for models with PM_{10} did not show large difference for BMA under either of the prior choices (Table 1).

Table 1 Summary of the posterior distribution of relative risk associated with a $20 \text{ ug}/m^3$ increase in all PM_{10} under BMA using ACAPS data set

Prior	Posterior mean of relative risk	95% posterior probability interval of relative risk
AIC	0.9936	(0.9861, 1.0085)
BIC	0.9913	(0.9033, 1.0987)
Local EB	0.9926	(0.9111, 1.0905)

2.5 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 based on a real data analysis in ACAPS described in the previous section.

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

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

$$\log(\mu_t) = \log(\mu_0) + \beta_0 \tilde{PM}_{10_lag0} + 0.25 * Trend + temp.s + \eta I_{DOW} . \quad (2.16)$$

μ_0 in (2.16) represents the mean of daily cardiopulmonary hospital admissions over the 6-year period and was estimated from ACAPS data as 115.07, β_0 is the true PM₁₀ effect, η are the true effects for day of the week and \tilde{PM}_{10_lag0} is the simulated PM₁₀ series that were created as followings. The effects of the PM₁₀ and day of the week were initially estimated by fitting the model

$$\begin{aligned} Y_t &\sim \text{Poisson}(\mu_t) \\ \log(\mu_t) &= \alpha_0 + \beta_0 PM_{10_lag0} + ns(time, df = 5 / year) \\ &\quad + ns(MNTP, df = 7) + \eta I_{DOW} \end{aligned} \quad (2.17)$$

to the observed ACAPS data and where Y_t is the count of daily cardiopulmonary hospital admissions, which follows a Poisson distribution with mean μ_t , β_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 PM₁₀, $ns(time, df = 5 / year)$ is the natural cubic spline function of calendar time with 5 degrees of freedom per year, $ns(MNTP, df = 7)$ 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 α_0 and η are unknown parameters.

The \tilde{PM}_{10_lag0} values in (2.16) were based on the following scheme. Since the degree of concurvity found in the ACAPS data was 0.613, we introduced this same degree of concurvity into the simulated data. The degree of concurvity in the ACAPS data was estimated by the correlation between the series of observed daily PM₁₀ (PM_{10_lag0}) and the corresponding fitted values (\hat{PM}_{10_lag0}) from an additive model. The additive model was

$PM_{10_lag0} = ns(time, df = 5 / year) + ns(MNTP, df = 7) + error$. For the simulation, a new PM_{10} series (\tilde{PM}_{10_lag0}) was generated by $\tilde{PM}_{10_lag0} = \hat{PM}_{10_lag0} + N(0, \sigma^2)$, where σ^2 was chosen so that the correlation between \tilde{PM}_{10_lag0} and \hat{PM}_{10_lag0} was equal to the estimated degree of concavity.

The other terms in (2.16), Trend and *temp.s*, are the effects for long-term and seasonal trend, and daily mean temperature, respectively.

The long-term trend and seasonality data for 6-year time series was generated using (Bateson and Schwartz 1999)

$$Trend = 1 + 0.6 * \cos(2\pi * \frac{day}{365.25}) + 0.4 * \cos(2\pi * \frac{day}{365.25}) * I(1358 < day < 1732).$$

The factor used to rescale the trend effect in (2.16) is set to be 0.25, rather than 0.2, as was used by Bateson and Schwartz (1999). Because a rescaling factor of 0.2 did not result in a satisfactory trend in our data, we tried a range of rescaling factors and found that a rescaling factor of 0.25 results in a trend effect similar to the real data. A comparison of the observed and simulated long-term and seasonal trend pattern in Figure 3 indicates the similarity of the pattern and the coherence of the peaks.

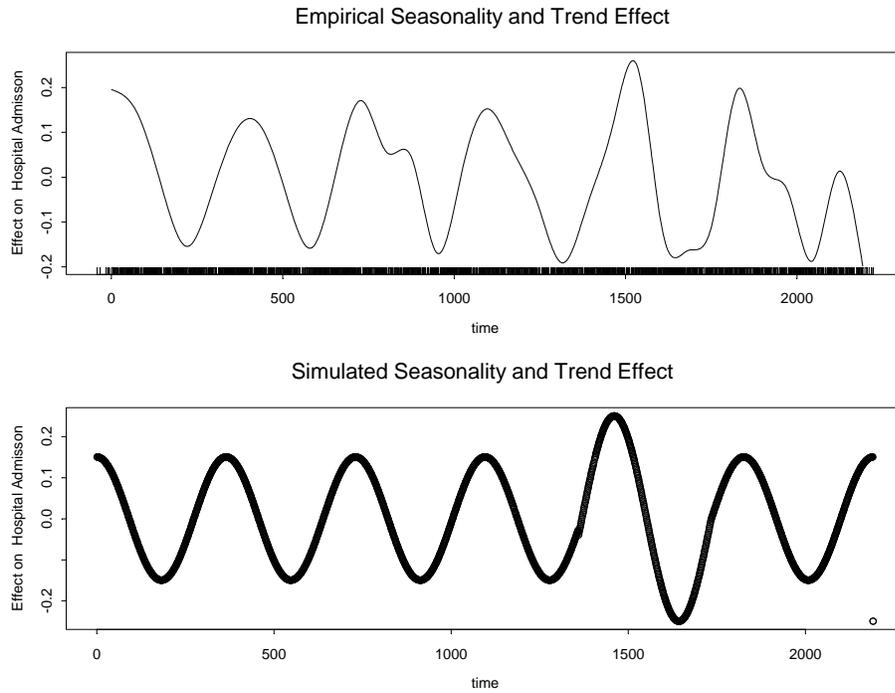


Figure 3 Empirical and simulated effect of seasonal and long-term trend on hospital admissions

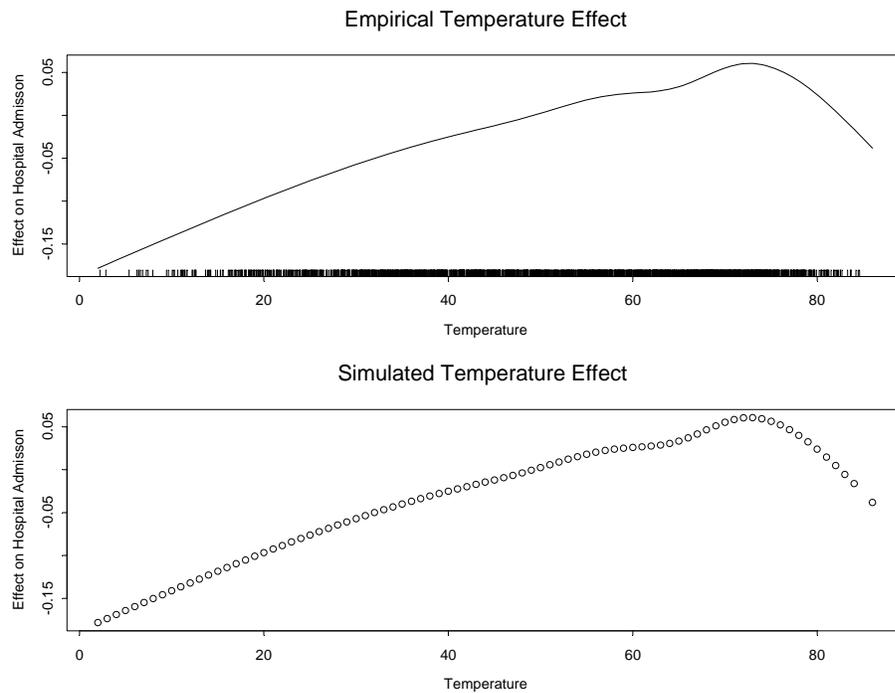


Figure 4 Empirical and simulated effect of temperature on hospital admissions

The daily mean temperature series, $temp.s$, was estimated from (2.17) by $temp.s = X_n \times beta.temp$, where X_n is a basis matrix generated from $ns(MNTP, df=7)$ in S-plus and $beta.temp$ is a vector of the estimated coefficients for temperature in (2.17). The comparison of the observed and simulated temperature pattern in Figure 4 indicates the similarity of the pattern.

We generated 1000 sets of 2190 observations for the hospital admissions. BMA analyses under AIC prior, BIC prior, and local EB estimates were conducted for all simulations and summary statistics were calculated. We selected a value similar to that estimated from the ACAPS data, 1.0003, to represent the true risk for a $20 \mu g/m^3$ increase in the same day level of PM_{10} . 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 of PM_{10} . Therefore, two model forms were used: one included all the predictor variables in (2.17), which only includes the same day PM_{10} term, and the other model added PM_{10} lag variables for the five previous days.

With the model that included only the same day level of PM_{10} the BMA method consistently and accurately selected the same day level of PM_{10} . The posterior means of relative risk were close to the true value of relative risk, which is greater than 1, for BMA under either of the prior choices (Table 2). With the model that contained same day level of PM_{10} and PM_{10} lagged by five days was used, the BMA approach correctly selected the models that have only the same day level of PM_{10} as the effect of air pollutant 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 of air pollutant. However, the posterior means of relative

risk had changed to be smaller than 1. The biased estimates for the relative risk may be because the inclusion of PM₁₀ lag variables increases the concurvity.

Table 2 Posterior means of relative risk associated with a 20 ug/m³ increase in all PM₁₀ and their 95% posterior probability intervals under BMA using simulated data set[†]

PM ₁₀ covariate in the fitted model	AIC	BIC	Local EB
Current day of PM ₁₀	1.0006 (0.9975, 1.0030)	1.0009 (0.9972, 1.0036)	1.0008 (0.9973 , 1.0034)
Current day and 5 previous days of PM ₁₀	0.9993 (0.8695,1.1495)	0.9984 (0.8468,1.1792)	0.9992 (0.8558 , 1.1677)

[†] The true relative risk for a 20 ug/m³ increase in PM₁₀ derived based on model (2.17) is given as 1.003

Our simulation results have indicated that when the observed pollution data follow model (2.16), the BMA method provides consistent estimates of the posterior mean of relative risk for models that incorrectly include lag terms for the pollution term, and these results hold for all of the three BMA implementations we have considered.

2.6 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 variable

selection procedure, which we investigated through the modeling of the ACAPS data set and (2) the uncertainty of the functional form of the degree of lagging for the ambient levels of a pollutant (an independent variable), which we investigated through the modeling simulated data based on the ACAPS data set.

For the variable selection analysis 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 0.991 and 0.994 for a $20 \mu\text{g}/\text{m}^3$ increase in all PM_{10} . While Arena *et al.* (2006) reported a higher risk of 1.226 for the current day level of PM_{10} , BMA method is more favorable due to the justification of uncertainties that have been ignored in model fitting. These results suggest that BMA is a useful method of reducing the uncertainty in the selection of model variables, and 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 the functional form of the independent variables, we found that the choice of the degree of lagging for the air pollution term was important. We found that if the lag variables of PM_{10} were incorrectly considered in the model, the posterior means of relative risk could change; in our case the change was one either side of the important relative risk of 1.0. We suspect that the shift might be due to the concurvity problem that results from the inclusion of lag variables of 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. We may, in future research, 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 in the future work. Note 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 optimum degree of lagging for variables, not based on only maximizing the likelihood, but by considering the possible effects of concavity, consistency of degree of lagging, and biological plausibility.

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. We suggest that research on this source will be interesting and informative.

2.7 ACKNOWLEDGEMENTS

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

**3.0 GENERALIZED LINEAR MIXED MODELS APPROACH IN TIME SERIES
STUDIES OF AIR POLLUTION**

Ya-Hsiu Chuang¹, Taeyoung Park², Gong Tang¹, Vincent C. Arena¹, Mark J. Nicolich³
and Sati Mazumdar¹

¹Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh,
Pennsylvania, U.S.A.

²Department of Statistics, School of Arts and Sciences, University of Pittsburgh,
Pennsylvania, U.S.A.

³ Lambertville, New Jersey, USA

3.1 ABSTRACT

Generalized additive model (GAM) with natural cubic splines (NS) has been commonly used as a standard analytical tool in time series studies of health effects of air pollution. Degrees of smoothing in the fit of GAM have been an issue with respect to the estimation of the health effects of air pollution. While studies indicated that larger degrees of smoothing than those estimated from data-driven methods should be specified and would result in less biased estimates on the effects of air pollution, oversmoothing the smooth functions can produce confounding bias and affect the true effects of air pollution. Instead of specifying fixed degrees of smoothing on the smooth functions, in this paper, degrees of smoothing are assumed to be random, which are from a common distribution in the generalized linear mixed modeling approach. We conducted a simulation study to assess the performances of generalized linear mixed model with natural cubic splines (GLMM + NS) with respect to the parameter estimates on the health effects of air pollution and their standard errors. Our simulation results showed that for smaller true effect of air pollution, GLMM + NS resulted in smaller empirical standard deviations of the estimates of the effect of air pollution than generalized linear model with natural cubic splines (GLM + NS), whereas the parameter estimates from GLM + NS were less biased than GLMM + NS. An illustrative example using data from Allegheny County Air Pollution Study (ACAPS) was given to compare the estimates of air pollution effects using GLMM + NS, GLM + NS and generalized additive model with smoothing splines (GAM + S).

Keywords: smoothing spline, natural cubic spline, degrees of smoothing, random effects

3.2 INTRODUCTION

Generalized additive models (GAMs) have been applied as a standard analytic tool in time series studies of air pollution for more than a decade (Schwartz 1994a, Clyde 2000, Dominici *et al.* 2002b). GAMs, which extend the application of generalized linear models (GLMs) by allowing non-parametric smoothers in addition to the parametric forms combined with a range of link functions, have the advantage of providing a good fit with the data when the non-linear associations between the outcomes and covariates exist. To adjust for the long-term trends and seasonality, as well as the short-term effects (e.g. temperature), GAMs include smooth functions of selected covariates to capture the shape of the relationship between covariates and the outcome. The degrees of smoothing are found to have more effects than the choices of smoothers on the estimation of the effects of air pollution (Rupert *et al.* 2003, Peng *et al.* 2006). Moreover, Peng *et al.* (2006) showed that the degrees of smoothing determined by data-driven methods, such as Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) that optimize the predictivity of the data series, may not give accurate estimates on the effects of air pollution, especially under high concavity which is common in much of the air pollution data. They focused on the smooth function of time and concluded that “the automatic use of criteria such as generalized cross-validation or AIC to select the degrees of freedom could be potentially misleading, particularly with high concavity, since they are designed to choose the degrees of freedom that will lead to optimum prediction of the mortality series, not necessarily to accurate estimation of β ” (where β is the effect of air pollution). Furthermore, the degrees of smoothing derived from these methods are assumed to be fixed over the study period. However, this assumption cannot be true if the smooth function is wiggly in some subsets of the study period and smooth in others.

In order to allow the degrees of smoothing to vary locally (e.g. vary within a short period of time, instead of the whole study period), we applied mixed models by assuming the terms that involve the knot locations of the smooth functions to be random. As a smooth function can be estimated by adapting the basis expansions, one can choose the dimension of the basis function with pre-determined degrees of freedom to achieve the flexible representations of a smooth function but penalize the basis coefficients for overfitting to ensure the smoothness of the functions. The smoothing methods are incorporated within the mixed models analytical frameworks by allowing the penalty to act differently for each spline basis, so that the fitted smooth functions could correctly capture the true functions. This can be achieved by constraining the basis coefficients to come from a common distribution. Thus, the local structure of the relationship between an outcome and covariates can be estimated more precisely.

As the smooth functions of covariates in GAMs are assumed to have random effects, the model becomes generalized additive mixed models (GAMMs). Moreover, GAMMs reduce to generalized linear mixed models (GLMMs) when the natural cubic splines are used.

In this paper, the connection between mixed models and semi-parametric regression models that achieve smoothing using natural cubic splines was elucidated, where the connection between mixed models and semi-parametric regression models has been explicitly depicted in Ruppert *et al.* (2003). We demonstrated how the relationship between an outcome and covariates can be modeled semi-parametrically within the framework of parametric mixed models. The quantity of interest is the effect of same day level of PM_{10} on cardiopulmonary hospital admissions. The performances of generalized linear mixed model with natural cubic splines (GLMM + NS) with respect to the bias and variance estimates of the effects of air pollution are evaluated by a simulation study. We compare the derived estimates to those from generalized

linear model with natural cubic splines (GLM + NS). While GLM + NS was found to outperform GAM + S with respect to the bias and variance estimates of the effect of air pollution when co-collinearity occurs (He *et al.* 2006), we compare the quantity of interest from models in GLMM + NS, GLM + NS, and generalized additive model with smoothing splines (GAM + S) in the illustrative example.

In Section 3.3, a review of spline smoothing for semi-parametric regression and GAMs is provided. The formulation of fitting GLMM with natural cubic splines is given in Section 3.4. A simulation study is presented in Section 3.5 for a comparative analysis using GLM + NS and GLMM + NS. Section 3.6 illustrates the application of the GLMM + NS modeling approach with Allegheny County Air Pollution Study (ACAPS) data (Arena *et al.* 2006), where ACAPS is a time-series study to investigate the effects of ambient air level of PM₁₀ on daily cardiopulmonary hospital admissions in elderly residents of Allegheny County from 1995 through 2000. Conclusions and discussions are given in Section 3.7.

3.3 REVIEW OF GENERALIZED ADDITIVE MODELS AND CHOICES OF SMOOTH FUNCTIONS

3.3.1 Generalized Additive Model (GAM)

A generalized additive model (GAM), defined as a generalized linear model with linear predictors involving a sum of smooth functions of covariates (Wood, 2006), is given by

$$y \sim \text{exponential family distribution}(\mu)$$

$$g(\mu) = X_*\beta_* + \sum_j f_j(T_j), \quad (3.1)$$

where y is a vector of an independent observed response variable with $\mu \equiv E(y)$, g is a link function, β_* is an unknown vector of the corresponding parameters, and $f_j(\cdot)$ are the smooth functions of covariates T_j . GAM allows examining the possible association between specific factors when non-linear relationships cannot be ruled out and controlling for potential non-linear covariates (Hastie and Tibshirani, 1990).

3.3.2 Choices of Smooth Functions

The fit of GAM requires the specification of the smooth functions in (3.1). The most common smooth functions in the study of air pollution include natural cubic splines (NS), smoothing splines (S) and penalized splines (P-splines).

Natural cubic splines are a type of regression splines with piecewise cubic splines joined at distinct knots with constraints that are linear beyond the boundary knots. Natural cubic splines with K knots can be expressed by the basis functions:

$$f(T) = B(T)\delta, \quad (3.2)$$

where $B(T) = [N_1(T) \ N_2(T) \ \cdots \ N_{k+2}(T) \ \cdots \ N_K(T)]$ with

$$N_1(T) = 1, N_2(T) = T, N_{k+2}(T) = d_k(T) - d_{k-1}(T),$$

$$d_k(T) = \frac{(T - c_k)_+^3 - (T - c_K)_+^3}{c_k - c_K}, \quad (T - c_k)_+^3 = (T - c_k)^3 \text{ if } T > c_k \text{ and } 0 \text{ otherwise, and } \delta \text{ is a}$$

vector of unknown parameters (Hastie *et al.* 2001). The choices of knot locations and the number of knots have substantial effects on the resulting smooth.

To avoid this sensitivity relating to the knots, both smoothing splines and penalized splines allow a large number of knots and then constrain their influence using the penalty terms to avoid overfitting. Smoothing splines are a natural spline fit but with knots at every data point. While smoothing splines would provide smoother functions than natural cubic splines, they could be computationally intensive.

Under a simplified case where $y = f(T) + \varepsilon$ with y as a vector of response variable and ε as a vector of random variables $\varepsilon \sim N(0, \sigma^2)$, penalized splines are defined as the smooth functions which minimize the penalized least squares (PLS)

$$\|y - f(T)\|^2 + \lambda \int [f''(T)]^2 dT, \quad (3.3)$$

where f can be represented by some basis functions such that

$$f(T) = B(T)\delta,$$

with $B(T)$ as a vector of spline basis functions and δ as a vector of unknown parameters. The first term in (3.3) measures the closeness to the data, the second term describes the roughness penalty and λ is a fixed smoothing parameter that controls the amount of smoothing. As $\lambda \rightarrow \infty$, the penalty term dominates, which forces $f''(T) = 0$ and leads to a straight line estimate for f . As $\lambda \rightarrow 0$, the penalty term becomes unimportant and results in an un-penalized regression spline estimate. As $f(T)$ is linear in the parameters δ , the penalty can be expressed as a quadratic form in δ , $\int [f''(T)]^2 dX = \delta^T S \delta$, where S is a symmetric positive semi-definite matrix of known coefficients. Hence, (3.3) can be re-written as:

$$\|y - f(T)\|^2 + \lambda \delta^T S \delta. \quad (3.4)$$

Penalized splines combine the most attractive attributes of regression splines and smoothing splines. They provide a more flexible way to model the non-linear relationships by

retaining a large number of knots in the regression splines formulation, but penalizing the piecewise polynomial coefficients for overfitting. The smoothness of the fit is controlled by a smoothing parameter, related to the severity of the constraint on the regression coefficients. The fit of penalized splines is found to be insensitive to the location of the knots as long as enough number of knots is specified.

3.3.3 Estimation of GAMs

The estimation of GAMs can be based on either the backfitting algorithm or the penalized likelihood maximization with integrated smoothness estimation via generalized cross validation (GCV) (Hastie and Tibshirani 1990, Wood 2006). The backfitting algorithm estimates each smooth component of an additive model by iteratively smoothing partial residuals from the additive models, with respect to the covariates to which the smooth functions relate. In this paper, as we are interested in relating GAMs to the mixed models, we focus on the penalized likelihood maximization. The idea of penalized least-squares in (3.4) can be generalized to the penalized likelihood.

In regression analysis, while the log-likelihood of model (3.1) can be used to find f_j , the likelihood can be maximized by any function that interpolates the data and the maximization of the log-likelihood cannot provide a sensible estimate of f_j . Good and Gaskins (1971) suggested subtracting from the log-likelihood a roughness penalty, which measures the local variation in f_j . This results in a penalized log-likelihood (Green and Silverman, 1994). The penalized likelihood can be written as

$$l_p(\beta, f_1, \dots, f_p) = l(\beta, f_1, \dots, f_p) - \frac{1}{2} \sum_{j=1}^p \lambda_j \int [f_j''(T_j)]^2 dT_j,$$

where $l(\beta, f_1, \dots, f_p)$ is the log-likelihood of model (3.1) and $\sum_{j=1}^p \lambda_j \int [f_j''(T_j)]^2 dT_j$ is the roughness penalty. Let the basis functions for each of the smooth functions be $f_j(T_j) = B_j(T_j)\delta_j$.

Then we have

$$l_p(\beta, f_1, \dots, f_p) = l(\beta, f_1, \dots, f_p) - \frac{1}{2} \sum_{j=1}^p \lambda_j \beta^T S_j \beta, \quad (3.5)$$

where $\beta^T = [\beta_*, \delta_1^T, \delta_2^T, \dots, \delta_p^T]$ and λ_j is the smoothing parameter for the smooth function $f_j(T_j)$. Maximization of (3.5) is equivalent to the minimization of (3.4).

The penalized log-likelihood in (3.5) can be maximized by using penalized iteratively re-weighted least squares (Wood, 2006). By defining $S = \sum_j \lambda_j S_j$ with λ_j is assumed to be known, $l_p(\beta, f_1, \dots, f_p)$ can be re-written as

$$l_p(\beta, f_1, \dots, f_p) = l(\beta, f_1, \dots, f_p) - \frac{1}{2} \beta^T S \beta. \quad (3.6)$$

The log likelihood for β and f_1, \dots, f_p for a generalized linear model is given by

$$l(\beta, f_1, \dots, f_p) = \sum_{i=1}^n \left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c_i(\phi, y_i) \right\},$$

where y follows a distribution in the exponential family with $f(y; \theta) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(\phi, y) \right\}$,

where $g(\mu) = X\beta$ and

$$E(y) = \mu = b'(\theta),$$

$$\text{Var}(y) = a(\phi) b''(\theta) = a(\phi) g^{-1}(\mu),$$

θ is a canonical parameter, ϕ is an arbitrary scale parameter and a , b and c are arbitrary functions. Let $a(\phi) = \phi/w$ with w as a known constant, $Var(y) = \phi b''(\theta)/w = \phi Var(\mu)$, where $Var(\mu) = b''(\theta)/w$. By chain rule,

$$\begin{aligned}\frac{\partial l}{\partial \beta_j} &= \frac{1}{\phi} \sum_{i=1}^n w_i \left[y_i \frac{\partial \theta_i}{\partial \beta_j} - b'(\theta_i) \frac{\partial \mu_i}{\partial \beta_j} \right] \\ &= \frac{1}{\phi} \sum_{i=1}^n \frac{(y_i - \mu_i)}{Var(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j}\end{aligned}$$

and (3.6) becomes

$$\frac{\partial l_p}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^n \frac{(y_i - \mu_i)}{Var(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} - [S\beta]_j. \quad (3.7)$$

Wood (2006) showed that the solutions for β are equivalent to those in solving the penalized non-linear least squares problem

$$S_p = \sum_{i=1}^n \frac{(y_i - \mu_i)^2}{Var(y_i)} + \beta^T S \beta.$$

By defining the pseudodata $z = g(\mu) + g'(\mu)(y - \mu)$,

$$S_p \approx \| \sqrt{W} (z - X\beta) \|^2 + \beta^T S \beta, \quad (3.8)$$

where $W = \{Var(\mu)[g'(\mu)^2]\}^{-1}$ is a diagonal weight matrix. The maximum penalized likelihood estimates, $\hat{\beta}$, can be obtained by the following iterative procedures. Let $\hat{\beta}^{[k]}$ represent the current estimates at the k^{th} iteration,

- 1) $z^{[k]}$ and $W^{[k]}$ can be calculated by $\eta^{[k]} = X\hat{\beta}^{[k]}$ and $\mu^{[k]} = g^{-1}(\eta^{[k]})$
- 2) Minimize (3.8) with respect to β to find $\hat{\beta}^{[k+1]}$, and $\eta^{[k+1]} = X\hat{\beta}^{[k+1]}$ and $\mu^{[k+1]}$. Increment k by one.

3) The PLS estimators are defined upon convergence.

The *gam* function in *mgcv* package in R uses penalized likelihood maximization method for estimation and can be used to fit the GAMs.

3.4 GENERALIZED LINEAR MIXED MODELS WITH NATURAL CUBIC SPLINES

As natural cubic splines can be expressed as parametric forms, we adapted (3.2) into (3.1) and a generalized additive model reduces to a generalized linear model which has the form of

$$\begin{aligned}
 y &\sim \text{exponential family}(\mu) \\
 g(\mu) &= X\beta + \sum_j f_j(T_j) = X\beta + \sum_j B(T_j)\delta_j \\
 &= X\beta + \{N_1(T_1)\delta_{11} + \dots + N_{K_1}(T_1)\delta_{1K_1}\} + \dots \\
 &\quad + \{N_1(T_j)\delta_{j1} + \dots + N_{K_j}(T_j)\delta_{jK_j}\}
 \end{aligned}$$

where g is a link function and $B_j(T_j)$ and δ_j are defined in Section 2. As the terms of $N_1(T_1)\delta_{11}, N_1(T_2)\delta_{21}, \dots, N_1(T_j)\delta_{j1}$ are constants, these terms could be combined with the intercept terms. In order to capture the local structure of the relationship between a response variable and covariates, we assume random coefficients on terms that involve the knot locations, where these coefficients are constrained to be from a common distribution. This results in a generalized linear mixed model with natural cubic splines (GLMM + NS) that is expressed by

$$\begin{aligned}
 y &\sim \text{exponential family}(\mu) \\
 g(\mu) &= X\beta + Zb, \tag{3.9}
 \end{aligned}$$

where X is an $n \times (m+1+j)$ fixed effects design matrix

$$X = [1 \quad X_1 \cdots X_m \quad N_2(T_1) \cdots N_2(T_j)]$$

with n as the number of observations and m as the number of covariates for the fixed effects, Z is an $n \times (q_1 + \cdots + q_j)$ random effects design matrix

$$Z = [Z_1 \quad \cdots \quad Z_q]$$

with $Z_j = [N_3(T_j) \quad \cdots \quad N_{K_j}(T_j)]$ as a $n \times q_j$ matrix with $q_j = K_j - 2$, where $N_{K_j}(T_j)$ is K_j^{th} basis function for covariate T_j defined in (3.2), K_j is the number of knots for covariate T_j , β is the vector of fixed effects parameters, and b is a $(q_1 + \cdots + q_j) \times 1$ vector of random effects parameters, where

$$b^T = [\delta_{13} \cdots \delta_{1K_1} \quad \cdots \quad \delta_{j3} \cdots \delta_{jK_j}].$$

Additionally, we assumed that $b \sim N(\tilde{b}, G)$ with $G = G(\theta)$, where θ is a $q \times 1$ unknown vector of variance components with q as the dimension of the random effects. It is also assumed that random effects are independent of each other.

To estimate the parameters in (3.9), the integrated quasi-likelihood is considered. We assumed that, given the random effects b , y 's are conditionally independent with means and variances specified as

$$E(y | b) = \mu = h(X\beta + Zb),$$

$$\text{Var}(y | b) = \phi \text{Var}(\mu),$$

where $g = h^{-1}$ and ϕ is the dispersion parameter for model (3.1), which is assumed to equal to 1 for the Poisson models. The integrated quasi-likelihood used to estimate (β, θ) is given by

$$L(\beta, \theta) = \frac{1}{\sqrt{(2\pi)^q |G(\theta)|}} \int \exp\left[-\frac{1}{2\phi} d(y; \mu) - \frac{1}{2} b^T G^{-1}(\theta) b\right] db,$$

where q is the dimension of the random effects b and $d(y; \mu) = -2 \int_y^\mu \frac{y-u}{\text{Var}(u)} du$. As the integral

does not have closed form solution, Laplace's method was applied to

$$PQL(\beta, b) = -\frac{1}{2\phi} d(y; \mu) - \frac{1}{2} b^T G^{-1}(\theta) b \quad (3.10)$$

for integral approximation (Breslow and Clayton 1993), which is referred to as the penalized quasi-likelihood (PQL). PQL is replaced by its quadratic expansion at $\hat{b} = \arg \min PQL(\beta, b)$ for fixed θ and β , and $\hat{\beta} = \arg \min PQL(\beta, \hat{b})$ for fixed θ . The approximations lead to the standard restricted maximum likelihood (REML) equations for θ by a working vector,

$$y^* = g(\hat{\mu}) + (y - \hat{\mu})g'(\hat{\mu}),$$

where $\hat{\mu} = g^{-1}(X\hat{\beta} + Z\hat{b})$. Thus, the distribution of y^* follows a linear model

$$y^* = X\hat{\beta} + Z\hat{b} + \varepsilon,$$

where $\varepsilon \sim N(0, W^{-1})$ and $W = \{\text{Var}(\mu)[g'(\mu)^2]\}^{-1}$.

The iterative procedures are summarized as follows:

- 1) Given θ and b , the fixed effects can be estimated by solving the normal equation

$$(X^T V^{-1} X)\beta = X^T V^{-1} y,$$

where $V = W^{-1} + ZG(\theta)Z^T$.

- 2) The random effects b can be estimated as $\hat{b} = G(\theta)Z^T V^{-1}(y - X\hat{\beta})$
- 3) Subsequently, the REML estimator for θ is

$$\hat{\theta}_j = \frac{\hat{b}_j^T \hat{b}_j}{q_j - t_{jj}}, \text{ for } j = 1, \dots, q,$$

where t_{jj} is the j^{th} diagonal element of $T = (I + Z^T SZD)^{-1}$ with $S = W - WX(X^T WX)^{-1} X^T W$.

- 4) Updates y^* at the end of each iteration and the PQL estimators are defined upon convergence (Harville 1977, Breslow and Clayton 1993).

The connection between mixed models and smoothing methods can be established by (3.7) and (3.10). As the penalized log likelihoods in (3.7) and (3.10) have the same form, we can maximize them to find the solutions for the estimates of (β, b) . The solutions for (3.10) are given by

$$\begin{bmatrix} \hat{\beta} \\ \hat{b} \end{bmatrix} = \arg \max_{\beta, b} \left\{ \log f(y|b) - \frac{1}{2} b^T G^{-1}(\theta) b \right\}, \quad (3.11)$$

where $\log f(y|b)$ is the log-likelihood; the solutions for (3.7) under natural cubic splines are given by

$$\begin{bmatrix} \hat{\beta} \\ \hat{b} \end{bmatrix} = \arg \max_{\beta, b} \left\{ l(\beta, b_1, b_2, \dots, b_j, f_1, \dots, f_p) - \frac{1}{2} \lambda^3 \begin{bmatrix} \beta \\ b \end{bmatrix}^T D \begin{bmatrix} \beta \\ b \end{bmatrix} \right\},$$

where $b^T = [b_1^T, b_2^T, \dots, b_j^T]$ and D is a known positive semi-definite penalty matrix. By assuming $D = [0, \dots, 0, 1, \dots, 1]$ where the elements corresponding to β are 0 and 1 for those corresponding to b , (3.7) is equivalent to (3.11). Let G_b be an identify matrix, the smoothing parameter λ can be derived by $\lambda^3 = 1/\sigma_b^2$. The degrees of smoothing (λ) are, hence, controlled by the variance components in the covariance matrix of b .

The PQL approach can be implemented by *glmmPQL* function in the *mass* package in R and the *GLIMMIX* procedure in SAS. We conducted our analysis in R.

3.5 SIMULATION STUDY

We conducted a simulation study to evaluate how the estimates of PM_{10} that have been found to be small in the study of air pollution vary with respect to the magnitude of the variance components of the random effects in GLMM + NS and GLM + NS. As GLM + NS performs better than GAM + S with respect to the bias and variance estimates when concavity exists in the data (He *et al.*, 2006), where we found 0.613 of degree of concavity in the real data from ACAPS, we do not consider the comparison with GAM + S in our simulation study.

The Poisson regression with random effects was used in the simulation study. For the Poisson regression, conditionally independent observations for the counts were generated from a GLMM + NS:

$$\log(\mu) = \beta_0 + \beta_{PM_{10_lag0}} * PM_{10_lag0} + \beta_{2,time} * time + \beta_{2,temp} * temp + \beta_{DOW} * I_{DOW} \\ + \sum_{j=3}^{df_{time}-2} b_{j-2,time} N_{j,time}(time) + \sum_{j=3}^{df_{temp}-2} b_{j-2,temp} N_{j,temp}(temp).$$

The covariates for level of PM_{10} , temperature, long-term trends and seasonality and day of the week were from the ACAPS study. The coefficients of fixed effects were set to be

$$\beta^T = (mean(\log(Y) - 3), \beta_{PM_{10_lag0}}, 0.1, -0.3, -0.01, 0.5, 0.5, 0.4, 0.4, 0.4),$$

where the coefficients were derived from the real data analysis using ACAPS data, $\beta_{PM_{10_lag0}}$ is the true health effect of same day level of PM_{10} on cardiopulmonary hospital admissions. The random effects of temperature and long-term trends and seasonality were generated from a multivariate normal distribution

$$\underline{b} \sim MVN\left(\begin{bmatrix} 0.1 \\ \vdots \\ 0.1 \end{bmatrix}, \begin{bmatrix} G_{temp} & \underline{0} \\ \underline{0} & G_{time} \end{bmatrix}\right),$$

$$\text{where } G_{temp} = \begin{bmatrix} \sigma_{temp}^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_{temp}^2 \end{bmatrix}_{(df_{temp}-2) \times (df_{temp}-2)} \quad \text{and} \quad G_{time} = \begin{bmatrix} \sigma_{time}^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_{time}^2 \end{bmatrix}_{(df_{time}-2) \times (df_{time}-2)} \quad \text{with}$$

degrees of freedom for temperature (df_{temp}) and seasonality (df_{time}) of 7 and 30, respectively.

These degrees of freedom of 7 for long-term trend and seasonality were the optimum degrees of freedom under AIC criteria and were determined by fitting the smooth function of this time-varying covariate with a range of degrees of freedom on cardiopulmonary hospital admissions using GLM + NS. After determining the degrees of freedom for long-term trend and seasonality, we then added six indicator variables for day of the week and the smooth function of temperature to the previous GLM + NS to account for the short-term effects. By considering a range of degrees of freedom for temperature in the fitted model, we selected the degrees of freedom for temperature by finding the model that has smallest AIC. We set $\beta_{PM_{10_lag0}}$ to have the true values ranging from 0.0001 to 0.001, σ_{temp}^2 with values in 0.01 and 0.05, and σ_{time}^2 with values in 0.01 and 0.10. 500 sets of 2190 observations for hospital admissions were generated.

Table 3 summarized the estimates of the coefficient of PM_{10_lag0} ($\hat{\beta}_{PM_{10_lag0}}$) defined as the mean of the coefficient estimates of PM_{10_lag0} over 500 runs, and their empirical standard deviations ($SD_{\hat{\beta}_{PM_{10_lag0}}}$) estimated from these coefficient estimates. The results indicated that while $\hat{\beta}_{PM_{10_lag0}}$ in GLMM + NS were more biased compared to those in GLM + NS, the empirical standard deviations of $\hat{\beta}_{PM_{10_lag0}}$ were found to be smaller in GLMM + NS than in GLM + NS.

Table 3 Empirical bias of $\hat{\beta}_{PM_{10_lag0}}$ with empirical standard deviations in the parentheses

True effect of PM ₁₀ ($\beta_{PM_{10_lag0}}$)	σ_{time}^2	σ_{temp}^2	<i>bias</i> *(<i>SD</i> _{$\hat{\beta}_{PM_{10_lag0}}$})	
			GLMM + NS	GLM + NS
0.001	0.1	0.05	6.50×10^{-5} (6.15×10^{-4})	6.00×10^{-6} (6.46×10^{-4})
		0.01	2.84×10^{-4} (5.63×10^{-4})	-1.10×10^{-5} (5.94×10^{-4})
	0.01	0.05	-1.00×10^{-4} (5.11×10^{-4})	-3.90×10^{-5} (5.55×10^{-4})
		0.01	1.11×10^{-4} (5.03×10^{-4})	0.00×10^{-5} (5.54×10^{-4})
0.0008	0.1	0.05	-8.40×10^{-5} (6.15×10^{-4})	-2.30×10^{-5} (6.44×10^{-4})
		0.01	2.01×10^{-4} (5.43×10^{-4})	-6.00×10^{-6} (5.78×10^{-4})
	0.01	0.05	-1.80×10^{-4} (5.79×10^{-4})	1.10×10^{-5} (6.07×10^{-4})
		0.01	-1.20×10^{-4} (6.03×10^{-4})	-1.60×10^{-5} (6.11×10^{-4})
0.0006	0.1	0.05	-7.40×10^{-5} (5.92×10^{-4})	-1.10×10^{-5} (6.20×10^{-4})
		0.01	1.95×10^{-4} (5.13×10^{-4})	-2.80×10^{-5} (5.45×10^{-4})
	0.01	0.05	6.80×10^{-5} (5.17×10^{-4})	-8.00×10^{-6} (5.54×10^{-4})
		0.01	2.04×10^{-4} (5.22×10^{-4})	0.00×10^{-6} (5.61×10^{-4})
0.0004	0.1	0.05	-7.20×10^{-5} (5.96×10^{-4})	-1.00×10^{-5} (6.21×10^{-4})
		0.01	-5.60×10^{-5} (4.80×10^{-4})	-2.00×10^{-6} (5.43×10^{-4})
	0.01	0.05	-6.90×10^{-5} (5.84×10^{-4})	1.10×10^{-5} (6.24×10^{-4})
		0.01	-8.50×10^{-5} (5.27×10^{-4})	-9.00×10^{-6} (5.89×10^{-4})
0.0003	0.1	0.05	-5.70×10^{-5} (6.15×10^{-4})	9.00×10^{-6} (6.46×10^{-4})
		0.01	2.40×10^{-4} (5.56×10^{-4})	2.40×10^{-5} (6.01×10^{-4})
	0.01	0.05	-1.50×10^{-4} (5.28×10^{-4})	1.10×10^{-5} (5.72×10^{-4})
		0.01	-1.78×10^{-4} (5.31×10^{-4})	3.00×10^{-6} (5.47×10^{-4})
0.0002	0.1	0.05	-6.70×10^{-5} (6.06×10^{-4})	-7.00×10^{-6} (6.36×10^{-4})
		0.01	2.23×10^{-4} (5.39×10^{-4})	1.70×10^{-5} (5.50×10^{-4})
	0.01	0.05	8.60×10^{-5} (5.35×10^{-4})	-8.00×10^{-6} (5.77×10^{-4})
		0.01	1.87×10^{-4} (5.38×10^{-4})	-2.40×10^{-5} (5.71×10^{-4})
0.0001	0.1	0.05	-8.10×10^{-5} (6.25×10^{-4})	1.60×10^{-5} (6.56×10^{-4})
		0.01	1.50×10^{-4} (5.58×10^{-4})	-7.00×10^{-6} (5.99×10^{-4})
	0.01	0.05	7.90×10^{-5} (4.94×10^{-4})	3.00×10^{-6} (5.31×10^{-4})
		0.01	1.24×10^{-4} (5.61×10^{-4})	6.00×10^{-6} (5.74×10^{-4})

The smaller estimated standard deviations for the estimates obtained from GLMM + NS compared to the corresponding values obtained from the model GLM+ NS were not expected and thus further examination of our methods for the generation of data and inferential procedures is needed. Firstly, the simulated data were generated under the model of GLMM + NS. Hence, it is possible that the estimates of the standard deviations of the coefficient estimates of PM_{10_lag0} in GLMM + NS have smaller values than those in GLM + NS. Robustness of the derived results should be judged by generating the data under GLM + NS or some other suitable curvilinear model that resembles the observed patterns of the data. Secondly, we should note that two

different functions in R were used for model fitting. The *glm* function in R, used to fit the model GLM + NS, computes the MLE of regression coefficients using the iteratively reweighted least squares method. The *glmmPQL* function in R, used to fit GLMM + NS model, computes the “approximate” MLE of regression coefficients using the penalized quasi-likelihood method with iterative procedures to approximate the likelihood function and hence, it does not compute the actual MLE. Moreover, PQL was found to give biased estimates of variance components (Breslow 2003). We should also note that the Laplace approximation used to fit GLMM in *glmmPQL* may be too simple. While *glmmPQL* function has the advantage of its speed and simplicity, alternative estimating approaches such as more accurate approximation of the penalized likelihood or MCMC need to be explored to derive less biased and more precise parameter estimates.

3.6 APPLICATION OF GENERALIZED LINEAR MIXED MODELS WITH NATURAL CUBIC SPLINES TO THE ACAPS DATA

3.6.1 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 (PM₁₀) 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 temperature data were used as the meteorological data in our study. Ambient air levels of a criteria pollutant (PM₁₀)

were recorded every hour for each of the 8 monitoring sites. The mean of the site-specific daily average PM_{10} values across all monitoring was used. Since only two sets of data out of 2192 were missing on dates 03/24/1998 and 11/04/1998, they were ignored and resulted in a total of 2190 observations for our data analysis.

3.6.2 Models Fitted

GLM + NS

Given ACAPS data, a GLM with natural cubic splines (GLM + NS) is given by:

$$\begin{aligned}
 y &\sim \text{Poisson}(\mu) \\
 \log[\mu] &= \alpha_0 + \beta_{PM_{10_lag0}} PM_{10_lag0} + ns(\text{time}, df_{\text{time}}) \\
 &\quad + ns(\text{temp}, df_{\text{temp}}) + \beta_{DOW} * I_{DOW} ,
 \end{aligned} \tag{3.12}$$

where $y = (y_1, \dots, y_n)^T$ is a vector of the observed counts of daily cardiopulmonary hospital admissions, which follows a Poisson distribution with mean μ_t , PM_{10_lag0} is the level of PM_{10} for the same day, $ns(\text{time}, df_{\text{time}})$ is the natural cubic splines function of calendar time with degrees of freedom df_{time} , $ns(\text{temp}, df_{\text{temp}})$ is the natural cubic splines function of temperature with degrees of freedom df_{temp} , and I_{DOW} are the six indicator variables for days of the week. The degrees of smoothing for long-term trends and seasonality were determined by fitting the smooth function of long-term trends and seasonality with a range of degrees of smoothing on cardiopulmonary hospital admissions using GLM + NS. Then, the optimum degrees of freedom were chosen from the fitted model which has the smallest AIC, where the smaller the AIC, the better the model fit. In addition, the residual plots were used to examine whether the seasonal

variation has been removed. We next considered the short-term effects by adding six indicator variables for day of the week and the smooth function of temperature into the previous generalized linear model, and repeated the same procedures that were used to find the degrees of smoothing for long-term trends and seasonality to get the degrees of smoothing for temperature. This resulted in 5 degrees of freedom per year for long-term trends and seasonality, and 7 degrees of freedom for daily mean temperature. Thus, the GLM + NS in (3.12) can be rewritten as

$$y \sim \text{Poisson}(\mu)$$

$$\log[\mu] = \alpha_0 + \beta_{PM_{10_lag0}} PM_{10_lag0} + ns(\text{time}, df = 5 / \text{year})$$

$$+ ns(\text{temp}, df = 7) + \beta_{DOW} * I_{DOW} .$$

As $ns(\text{time}, df = 5 / \text{year})$ and $ns(\text{temp}, df = 7)$ can be expressed by the basis functions defined in (3.2). This leads to

$$\log[\mu] = \beta_0 + \beta_{PM_{10_lag0}} * PM_{10_lag0} + \beta_{2,time} * \text{time} + \beta_{2,temp} * \text{temp} + \beta_{DOW} * I_{DOW}$$

$$+ \sum_{j=1}^{28} \beta_{j+2,time} N_{j,time}(\text{time}) + \sum_{j=1}^5 \beta_{j+2,temp} N_{j,temp}(\text{temp})$$

$$= X \beta , \tag{3.13}$$

where

- $X = [1 \quad PM_{10_lag0} \quad \text{time} \quad \text{temp} \quad I_{DOW} \quad N_3(\text{time}) \cdots N_{30}(\text{time}) \quad N_3(\text{temp}) \cdots N_7(\text{temp})]$ is an $n \times 43$ design matrix for the fixed effects with $N_j(\text{time})$ and $N_j(\text{temp})$ defined in (3.2)
- $\beta = [\beta_0 \quad \beta_{PM_{10_lag0}} \quad \beta_{2,time} \quad \beta_{2,temp} \quad \beta_{DOW} \quad \beta_{3,time} \cdots \beta_{30,time} \quad \beta_{3,temp} \cdots \beta_{7,temp}]$ is a 43×1 vector of the fixed-effect coefficients.

GAM + S

When the natural cubic splines in (3.12) are replaced by the smoothing splines, it results in a GAM with smoothing splines (GAM + S)

$$\begin{aligned}
 Y &\sim \text{Poisson}(\mu) \\
 \log[\mu] &= \alpha_0 + \beta_0 PM_{10_lag0} + s(\text{time}, df = 5 / \text{year}) \\
 &\quad + s(\text{temp}, df = 7) + \beta_{DOW} * I_{DOW} .
 \end{aligned} \tag{3.14}$$

GLMM + NS

To allow the degree of smoothing to vary locally in model (3.13), we incorporated the random effects into GLM. This resulted in GLMM with natural cubic splines (GLMM + NS)

$$\begin{aligned}
 \log[\mu] &= \beta_0^* + \beta_{PM_{10_lag0}} * PM_{10_lag0} + \beta_{1,time} * \text{time} + \beta_{1,temp} * \text{temp} + \beta_{DOW} * I_{DOW} \\
 &\quad + \sum_{j=1}^{28} b_{j+2,time} N_{j,time}(\text{time}) + \sum_{j=1}^5 b_{j+2,temp} N_{j,temp}(\text{temp}) \\
 &= X\beta + Zb,
 \end{aligned} \tag{3.15}$$

where

- $X = [1 \quad PM_{10_lag0} \quad \text{time} \quad \text{temp} \quad I_{DOW}]$ is an $n \times 10$ design matrix for the fixed effects
- $\beta = [\beta_0 \quad \beta_{PM_{10_lag0}} \quad \beta_{1,time} \quad \beta_{1,temp} \quad \beta_{DOW}]^T$ is a 10×1 vector of the fixed-effect coefficients
- $Z = [N_3(\text{time}) \cdots N_{30}(\text{time}) \quad N_3(\text{temp}) \cdots N_7(\text{temp})]$ is an $n \times 33$ design matrix for the random effects
- $b = [b_{3,time} \cdots b_{30,time} \quad b_{3,temp} \cdots b_{7,temp}]^T$ is a 33×1 vector of the random-effect coefficients with $b \sim N(0, G)$, where $G = \sigma_b^2 G_b$. G_b are assumed to have identity variance-covariance structures for the long-term trends and seasonality and temperature.

Table 4 summarizes the estimates for the effect of same day level of PM₁₀ (PM_{10_lag0}) and the corresponding standard errors for GLMM + NS, GLM + NS and GAM+ S. Since He *et al.* (2006) showed that GLM + NS outperforms GAM + S as concurvity increases in the data where we found medium-to-high degree of concurvity (0.61) in ACAPS data, we concluded that GAM + S would give a biased estimate on the effect of same day level of PM₁₀. The estimates of the health effects of same day level of PM₁₀ were found to be larger in GLMM + NS than GLM + NS, as well as the standard errors of the corresponding estimates. As GLMM + NS assumes the degrees of smoothing to be random, GLMM + NS were presumably to have larger standard errors than GLM + NS.

The variances of the random spline coefficients for temperature and long-term trends and seasonality are $\hat{\sigma}_{temp}^2 = 0.0199$ and $\hat{\sigma}_{time}^2 = 0.0492$. As a result, the smoothing parameters for temperature and seasonality and trend effects were estimated as $\hat{\lambda}_{temp} = 3.69$ and $\hat{\lambda}_{time} = 2.73$.

Root mean squared errors (RMSE) for goodness-of-fit, defined as $RMSE = \sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / n}$, were shown in Table 4. It is known that the methods used to determine the fixed degrees of freedom in GLM + NS and GAM + S were to optimize the predictivity of the data series, rather than to find the accurate estimates of PM_{10_lag0}. As a result, these two models consistently gave smaller RMSE than GLMM + NS.

Table 4 Summary for the fixed effect estimate of PM_{10_lag0}		
	$\beta_{PM_{10_lag0}}$ (SE*)	RMSE
GLM + NS	0.000167 (0.000163)	14.73
GAM + S	0.000377 (0.000150)	14.57
GLMM + NS	0.000277 (0.000222)	14.76

*Standard errors of the PM_{10_lag0} estimate

3.7 CONCLUSION AND DISCUSSION

In this study, we proposed a GLMM + NS to handle the problem relating to degrees of smoothing. The conventional data-driven methods that optimize the predictivity of the data series to determine the degrees of freedom were found to give biased estimates (Peng *et al.* 2006). While larger degrees of freedom than those derived from the optimization of the prediction of the data series may give a more accurate estimation of the effects of air pollution under high concavity, oversmoothing the smooth functions could produce confounding bias and affect the estimation of air pollution effects. Rather than assuming fixed degrees of freedom for the smooth functions over the whole study period, our method allows the degrees of smoothing to vary in its own way by assuming random effects on terms of smooth functions that related to the pre-specified knot locations.

In our simulation study, we found smaller standard deviations of the parameter estimates of the health effects of air pollution in GLMM + NS than GLM + NS. Intuitively, the standard deviations from GLMM + NS are supposed to be larger than GLM + NS. Results from our simulation studies indicate that more in depth analyses with special attention to inferential procedures using readily available software are needed to have any definitive conclusion about the performance of our proposed approach.

4.0 CONCLUSION AND DISCUSSION

The purpose of this dissertation is to provide methods to handle the current statistical issues in the time-series studies of air pollution and improve the estimates of the health effects of air pollution. We are able to account for the uncertainties resulting from the variable selection procedures that have commonly not been accounted for and may have resulted in incorrect conclusions, particularly in determination of the lagged effects of air pollution by adapting BMA. Furthermore, we addressed the issue regarding the degrees of smoothing by our proposed method, GLMM + NS. Our proposed model, GLMM + NS, was found to produce more precise estimates of the health effects of current day level of PM_{10} than the commonly used generalized linear models with natural cubic splines (GLM + NS) in our simulation studies. However, due to the limitations of the readily available software that used to derive the parameter estimates, further investigations are needed to have any definitive conclusion about the performance of our proposed model.

While BMA and GLMM + NS provide better ways to handle the issues relating to the selection of the optimum degree of lagging for variables and degrees of smoothing, some issues have not been considered in this dissertation and were discussed in Sections 2.6 and 3.7. Future research can be conducted that:

- Incorporates the uncertainties associated with the functional forms of the models in BMA
- Examines the performance of the estimations of the health effects of air pollution under different degrees of concavity in GLMM + NS

- Incorporates multiple lagged effects of air pollution into GLMM + NS

BIBLIOGRAPHY

- Arena VC, Mazumdar S, Zborowski JV, Talbott EO, He S, Chuang Y, Schwerha JJ. 2006. A Retrospective Investigation of PM₁₀ in Ambient Air and Cardiopulmonary Hospital Admissions in Allegheny County, Pennsylvania: 1995-2000. *JOEM* **48(1)**: 38-47
- Breslow NE, Clayton DG. 1993. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* **88**: 9-25
- Breslow, N. 2003. Whither PQL? *UW Biostatistics Working Paper Series Working Paper* 192
- Clyde M. 2000. Model Uncertainty and Health Effect Studies for Particulate Matter. *Environmetrics* **11**: 745-763
- Clyde M, George EI. 2004. Model Uncertainty. *Statistical Science* **19(1)**: 81-94
- Dominici F, McDermott A, Zeger SL, Samet JM. 2002b. On the Use of Generalized Additive Mmodels in Time-Series Studies of Air Pollution and Health. *American Journal of Epidemiology* **156**: 193-203
- Furnival GM, Wilson RW Jr. 1974. Regression by Leaps and Bounds. *Technometrics* **16**: 499-511
- George EI, Foster DP. 2000. Calibration and Empirical Bayes Variable Selection. *Biometrika* **87(4)**:731-747
- Good LJ, Gaskins.RA. 1971. Non-Parametric Roughness Penalties for Probability Densities. *Biometrika* **58**: 255-277
- Green PJ. 1987. Penalized Likelihood for General Semi-Parametric Regression Models. *International Statistical Review* **55**: 245-259
- Green PJ, Silverman BW. 1994. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Volume 58, Mongraphs on Statistics and Applied Probability, Chapman and Hall
- Hansen MH, Yu B. 2003. Minimum Description Length Model Selection Criteria for Generalized Linear Models. *IMS Lecture Notes* 145-163
- Harville DA. 1977. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of American Statistical Association* **72**:320:340

- Hastie T, Tibshirani R. 1990. Generalized Additive Models. London: Chapman & Hall
- Hastie T, Tibshirani R, Friedman J. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer
- He S, Mazumdar S, Arena VC. 2006. A Comparative Study of the Use of GAM and GLM in Air Pollution Research. *Environmetrics* **17**(1): 81-93
- Hoeting JA, Madigan D, Raftery AE, Volinsky CT. 1999. Bayesian Model Averaging: A Tutorial (with Discussion). *Statistical Science* **14**: 382-401
- Kass RE, Raftery AE. 1995. Bayes Factor. *Journal of the American Statistical Association* **90**(430): 773-795
- Madigan D, Raftery AE. 1994. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occams Window. *Journal of the American Statistical Association* **89**(428): 1535-1546
- Peng R, Dominici F, Louis T. 2006. Model Choice in Multi-Site Time Series Studies of Air Pollution and Mortality *Journal of the Royal Statistical Society Series A with discussion* **169**(2): 179-203
- Raftery AE. 1996. Approximate Bayes Factors and Accounting For Model Uncertainty in Generalised Linear Models. *Biometrika* **83**(2): 251-266
- Ruppert D, Wand MP, Carroll J. 2003. Semiparametric Regression. Cambridge University Press
- Schwartz J. 1993. Air Pollution and Daily Mortality in Birmingham, Alabama. *Am J Epidemiol* **137**(10): 1136-47
- Schwartz, J. (1994a) Nonparametric Smoothing in the Analysis of Air Pollution and Respiratory Illness. *Canadian Journal of Statistics* **22**: 471-488
- Smith RL, Davis JM. 2000. Regression Models for Air Pollution and Daily Mortality: Analysis of Data from Birmingham, Alabama. *Environmetrics* **11**(6): 719-743
- Taplin RH. 1993. Robust Likelihood Calculation for Time Series. *Aann Statist* **6**:461-464
- Taplin RH, Raftery AE. 1994. Analysis of Agricultural Field Ttrials in the Presence of Outliers and Fertility Jumps. *Biometrics*, **50**: 764-781
- Tierney L, Kadane JB. 1986. Accurate Approximation for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association* **81**: 82-86
- Volinsky CT, Madigan D, Raftery AE, Kronmal RA. 1997. Bayesian Model Averaging in Proportional Hazard Models: Assessing Stroke Risk. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **46**: 433-448
- Wood S. 2006. Generalized Additive Models: An Introduction with R. New York: Chapman & Hall

Wordley J, Walters S. 1997. Short Term Variations in Hospital Admissions and Mortality and Particulate Air Pollution. *Occup Environ Med* **54**(2): 108-16

Zellner A. 1986. On Assessing Prior Distributions and Bayesian Regression Analysis Using g -prior Distributions. In *Bayesian Inference and Decision Techniques—Essays in Honour of Bruno de Finetti*, Goel PK, Zellner A (eds). North-Holland: Amsterdam; 233–243