Model Uncertainty and Health Effect Studies for Particulate Matter

Merlise Clyde

Institute of Statistics and Decision Sciences
210A Old Chemistry Box 90251
Duke University
Durham, NC, 27708-0251

email: clyde@isds.duke.edu
phone: (919) 681–8440

October 12, 1999

This work was supported by NSF grants DMS-96.26135 and DMS-97.33013, and was completed while the author was a visitor at the National Research Center for Statistics and the Environment, University of Washington. Although the research described in this article has been funded in part by the United States Environmental Protection Agency through agreement CR825173-01-0 to the University of Washington, it has as not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

Model Uncertainty and Health Effect Studies for Particulate Matter

Summary

There are many aspects of model choice that are involved in health effect studies of particulate matter and other pollutants. Some of these choices concern which pollutants and confounding variables should be included in the model, what type of lag structure for the covariates should be used, which interactions need to be considered, and how to model nonlinear trends. Because of the large number of potential variables, model selection is often used to find a parsimonious model. Different model selection strategies may lead to very different models and conclusions for the same set of data. As variable selection may involve numerous tests of hypotheses, the resulting significance levels may be called into question, and there is the concern that the positive associations are a result of multiple testing. Bayesian model averaging is an alternative that can be used to combine inferences from multiple models and incorporate model uncertainty. This paper presents objective prior distributions for Bayesian model averaging in generalized linear models so that Bayesian model selection corresponds to standard methods of model selection, such as the Akaike Information Criterion (AIC) or Bayes Information Criterion (BIC), and inferences within a model are based on standard maximum likelihood estimation. These methods allow non-Bayesians to describe the level of uncertainty due to model selection, and can be used to combine inferences by averaging over a wider class of models using readily available summary statistics from standard model fitting programs. Using Bayesian Model Averaging and objective prior distributions, we re-analyze data from Birmingham, AL and illustrate the role of model uncertainty in inferences about the effect of particulate matter on elderly mortality.

KEYWORDS: AIC; BIC; Bayesian Model Averaging; Jeffreys' Prior; Model Selection; Noninformative Priors; Poisson Regression; PM10; RIC

1 Introduction

Statistical analyses of the effect of air pollution on human mortality have been performed for a multitude of cities and a number of authors have found statistically significant relationships between increased mortality in the elderly population (and other health outcomes) and increases in particulate matter (PM). Partly on the basis of epidemiological studies using PM_{10} (particulate matter with aerodynamic diameter less than 10 microns), the U.S. Environmental Protection Agency (EPA) in 1997 introduced new standards for $PM_{2.5}$ (particulate matter with aerodynamic diameter less than 2.5 microns). Scientists and policymakers recognized that that additional research was needed, and in 1998, the U.S. Congress directed the EPA to arrange for an independent study by the National Research Council on priorities for particulate matter research.

One concern raised by the 1998 National Research Council report (in Research Topic 10) is whether the positive associations between particulate matter and mortality (or other health outcomes) are an artifact of model selection due to multiple hypothesis testing. There are many aspects of model choice that are involved in health effect studies of particulate matter and other co-pollutants. Some of these choices concern which pollutants and confounding variables should be included in the model, what type of lag structure for the covariates should be used, which interactions need to be considered, and what adjustments should be made in the multiple time-series for long-term trends and seasonality. Generalized additive models (GAMs) for Poisson data are often used to model daily mortality, adjusting for nonlinear trends using smoothing splines or other semi-parametric approaches, and including smoothed functions and lags of meteorological variables and pollution variables.

From a classical perspective it may not be practical or desirable to include every conceivable confounding variable. As the number of variables or basis functions approaches the sample size, over-fitting becomes a serious concern and standard errors may be inflated, making it more difficult to detect real effects. Model selection is often used to find a parsimonious model that adjusts for the important confounding variables, while attempting to eliminate autocorrelation and overdispersion. This is often done in a highly exploratory fashion, and different model selection strategies may lead to different models and conclusions about the magnitude of relative risks associated with changes in particulate matter. For example, in analyses for Birmingham, AL, Schwartz (1993) found the best model had a relative risk of 1.11, based on 100 $\mu g/m^3$ increase in PM_{10}, using a PM_{10} exposure measure based on the average of PM_{10} for the three previous days. Smith *et al.* (1999) (to be referred to as SDSSS) found that results were sensitive to the exposure measure of PM_{10} used in the model. Using the same exposure measure as Schwartz, SDSSS found a statistically significant relationship between PM_{10} and non-accidental elderly mortality, but when allowing for model selection to determine which lags of PM_{10} to include, found that there was no significant effect.

For making inferences, the selected "best" model is often treated as if it were the true model. This procedure ignores the uncertainty involved in model selection, and may lead to overconfident predictions and policy decisions that are riskier than one thinks they are (Draper 1995, Hodges 1987). One may also find "significant" spurious effects, while the meaning of reported significance levels for the "best" model is also questionable (Viallefont *et al.* 1998). Model uncertainty often outweighs other sources of uncertainty (Hoeting *et al.* 1999), but is typically ignored in practice.

Bayesian Model Averaging (BMA) using hierarchical models provides a coherent approach for combining predictions and inferences from multiple models, and often leads to improved predictive performance and reduced frequentist risk (Clyde and George 1998, Lamon and Clyde 1998, Hoeting *et al.* 1999). With BMA, predictions and inferences are based on a set of models rather than a single model. Predictions are obtained by forming a weighted average of predictions over the different models, where the weights depend on the degree to which the data support each model. All variables are used, but coefficients for variables that are less important are shrunk towards zero. For the health effects models, model averaging can be used to incorporate uncertainty about which of several plausible pollution and meteorological variables are related to mortality, incorporate uncertainty about lag structures in pollution and meteorological variables, and can be used to account for uncertainty in the number of knots and locations of knots used in semi-parametric models for nonlinear trends.

While model averaging is straightforward to implement in theory, model averaging requires specification of prior distributions for parameters within models and prior weights for each model. While subjective prior distributions can be used to incorporate previous knowledge about health effects of particulate matter, this can be controversial because of questions of whose prior beliefs are being represented. In this situation, it may be desirable to have reference analyses based on "non-informative" prior distributions to supplement analyses based on subjective prior distributions. Also because of the complexity of generalized additive models and the number of confounding variables, even carefully elicited prior distributions may have unforeseen consequences on model selection. As part of an overall sensitivity analyses, objective prior distributions play an important role. By incorporating model uncertainty and considering a range of objective prior distributions, we can, perhaps, increase our confidence that positive associations are not an artifact of model selection.

In this paper, data from Birmingham, AL are re-analyzed using Bayesian Model Averaging (BMA) in conjunction with generalized additive models to assess the impact of model uncertainty on estimates of relative risks due to changes in PM_{10}. As in Schwartz (1993) and Smith *et al.* (1999), the response variable is non-accidental mortality. Additional information on the data and variables is given in section 2. In section 3, we describe the hierarchical Poisson regression model for model averaging. In section 4, we present a class of objective prior distributions for generalized linear models (GLMs) so that 1) inferences within a model are based on maximum likelihood theory, and 2) Bayesian model selection can be calibrated to standard methods of model selection, such as AIC (Akaike Information Criterion; Akaike 1973, 1978), BIC (Bayes Information Criterion; Schwarz 1978), RIC (Risk Inflation Criterion; Foster and George 1994) and other methods, through the choice of a single hyperparameter c. Such methods can be implemented using readily available summary statistics from standard model fitting programs. Model averaging using these objective prior distributions provides a bridge between classical and Bayesian methods of estimation, and presents a natural framework so that non-Bayesians can make inferences from multiple models via model averaging. In Section 5, we apply Bayesian model averaging with several objective prior distributions to the Birmingham data. We construct posterior distributions for relative risks based on a 100 $\mu g/m^3$ increase in PM_{10} for comparison with Schwartz (1993). These distributions incorporate model uncertainty as well as parameter uncertainty. We also conduct a small model validation study to compare model selection and model averaging under BIC and AIC prior distributions.

2 Variables

The data used in this analysis are the same as those used in SDSSS and are based on daily measurements from 1985–1988 of mortality (from the National Center for Health Statistics), PM_{10} (from the U.S. Environmental Protection Agency, EPA), and meteorology variables (from the U.S. National Climatic Data Center in Ashville, NC). Variable names and descriptions are given in Table 1.

[Table 1 here]

The response variable for this analysis is daily elderly non-accidental mortality, which is defined as the total number of deaths on a given day of all individuals age 65 and older, excluding deaths attributed to accidental causes. While Schwartz (1993) used total non-accidental mortality, the number of non-accidental deaths in individuals under 65 averages around 2 per day, and should not impact conclusions greatly.

PM_{10} data are available from the EPA's aerometric data base for 13 monitors in 8 locations in Jefferson County, AL, which contains the metropolitan area of Birmingham (Figure 1). These consist of readings from a daily monitor located in Birmingham (monitor ID 0023), a daily monitor in Leeds (monitor ID 1010), and several monitors (ID's 0002, 0012, 0026, 2003, 3003, 6002) throughout the county that collected data every 6 days. It appears that Schwartz (1993) used the daily mean of PM_{10} data from all available PM_{10} monitors within the metropolitan area to construct a daily area-wide measure of PM_{10}. Because the daily monitor in Leeds collected data only for the latter half of the time period and exhibited lower values on average than the daily monitor in Birmingham (Figure 2), the daily area-wide average varies substantially depending on whether the Leeds data are included. To avoid a bias in the area-wide average because the Leeds data are unavailable during the first half of the time period, we used PM_{10} data from the daily monitor within Birmingham (Monitor ID 0023), which started operation in August, 1985 and was in operation until the end of 1988. We will explore the difference in results based on using the area-wide average (`pma`) versus a single daily monitor 0023 (`pm`). Uncertainty in which monitors are representative of population exposure is an important question and an open area for research.

[Figures 1 and 2 here]

After exploratory modeling, Schwartz (1993) reported that the average of PM_{10} from the three previous days was the best predictor of mortality. SDSSS investigated using individual lags in addition to the 3-day average of PM_{10}. In order to take into account uncertainty in the lag structure, we constructed up to three day lags of PM_{10}, where lag 0 is the current day, and used these as separate variables in modelling.

The meteorological variables and lags (up to 2 days) are summarized in Table 1. As most papers include temperature, some measure of humidity, and sometimes atmospheric pressure, we include all of these in our list of potential variables. As low wind speeds are confounded with high pollution events, but are not expected to be associated with mortality (Schwartz 1993), daily average wind speed was eliminated from consideration.

The mortality data have a strong annual component that may be related to flu epidemics or other causes and longer term changes in population size. As long term trends and yearly variation in the covariates are confounded with long term and yearly variation in mortality, it is necessary to remove this source of variation from the analysis prior to assessing the effect of PM_{10}. This is often done by adding a smooth function of time to the model using smoothing splines (Smith *et al.* 1997) or sine-cosine functions (Schwartz 1993). The Time Series Working Group at the NRCSE Particulate Methodology Workshop held in Seattle, Oct 19-22,1998 (URL `http://www.nrcse.washington.edu/events/pm-workshop.html`) concluded " While the method used to remove longer-term variation is unlikely to be important, the choice of which time scales to include and which to exclude may influence the results. Removing too little information exposes the analysis to confounding by season, removing too much reduces the power of the analysis and may exclude important health effects." Most authors have used splines with roughly 5 to 12 knots per year to remove large scale variation that is thought not to be attributable to weather or particles.

We have found that using cubic splines, B-splines or thin-plate splines lead to little difference in results, but the number of knots does have a large impact on smoothness of the unknown trend. In what follows we report results using a thin-plate spline basis. To construct this basis, we selected 30 knots at equally spaced time points over the length of the sampling period, corresponding to roughly one knot per month. Let k_j denote a knot at location j, where $0 < k_j < n$. The jth basis element evaluated at the point t is constructed as

$$b_j(t) = (t - k_j)^2 \log(|t - k_j|).$$

A function $f(t)$, where t is the time index, representing the unknown trend can be represented as

$$f(t) = \alpha_0 + \sum_{j=0}^{K} \alpha_j b_j(t).$$

Removing knot j is equivalent to setting an α_j to zero. As the dimension or number of knots is unknown, this is also a variable selection problem; model uncertainty in the number of knots and locations can be addressed using BMA. For more discussion of Bayesian approaches to function estimation using thin-plate or other radial bases see Holmes and Mallick (1997).

3 Hierarchical Poisson Regression Model

Daily mortality reflects counts which are usually modeled with a Poisson or over-dispersed Poisson distribution. We will let Y_i denote the non-accidental elderly mortality for day i, $i = 1, \ldots, n = 1247$ and let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]$ denote the $(n \times p)$ design matrix based on all variables under consideration. The design matrix $\mathbf{X}$ can include basis terms for smoothing splines to model nonlinear trends, meteorological variables, such as temperature and humidity, pollution variables, lags of meteorological and pollution variables, seasonal indicators, or any other known confounders. We will focus on the set of variables listed in Table 1.

Under the full model, we assume that the observations $\mathbf{Y} = (Y_1, Y_2, \ldots Y_n)'$ are independent Poisson random variables with means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ and that the means are related to the covariates via a link function g,

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$$

and use the canonical log link function. For approaches using an identity link see Clyde (1999). In the present context of variable selection, models correspond to different probability specifications for the data, so that under the m^{th} model ($\mathcal{M}_m$)

$$\mathcal{M}_m : \quad \mathbf{Y} \sim \text{Poisson}(\boldsymbol{\mu}) \qquad \log(\boldsymbol{\mu}) = \mathbf{X}_m \boldsymbol{\beta}_m$$

where $\mathbf{X}_m$ is the design matrix under model $\mathcal{M}_m$ and $\boldsymbol{\beta}_m$ is the vector of regression coefficients for model $\mathcal{M}_m$. The set of possible models is given by $\mathcal{S} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_M\}$ and includes all subsets of potential covariates, or $M = 2^p$ models.

3.1 Hierarchical Model

Model uncertainty can be formally accounted for by building an expanded model that encompasses all models in $\mathcal{S}$. In constructing the hierarchical model, it is convenient to introduce a p dimensional vector $\boldsymbol{\gamma}$ of indicator variables where γ_j equals one if variable $\mathbf{x}_j$ is included under $\mathcal{M}_m$. To link $\mathcal{M}_m$ and $\boldsymbol{\gamma}$, we take $\boldsymbol{\gamma}$ to be the binary representation of m.

The hierarchical model is defined in three stages. The first stage of the hierarchy defines the distribution for the data in terms of all variables:

$$\mathbf{Y}|\boldsymbol{\beta}, \mathcal{M}_m \sim \text{Poisson}(\exp(\mathbf{X}\boldsymbol{\beta})) \tag{1}$$

with density $f(\mathbf{Y}|\boldsymbol{\beta}, \mathcal{M}_m) = f(\mathbf{Y}|\boldsymbol{\beta})$. In the expanded model, variables are eliminated by allowing coefficients to be exactly 0. This is achieved by allowing point masses at zero in the prior distribution for $\boldsymbol{\beta}$ given $\mathcal{M}_m$ in the second stage. Coefficients for variables not in $\mathcal{M}_m$ are identically zero as specified through distributions $\delta_0(\beta_j)$ that are degenerate at zero, while coefficients for variables included under $\mathcal{M}_m$ ($\boldsymbol{\beta}_m$) have non-degenerate prior distributions, so that the joint distribution for $\boldsymbol{\beta}$ given $\mathcal{M}_m$ is

$$p(\boldsymbol{\beta}|\mathcal{M}_m) = p(\boldsymbol{\beta}_m|\mathcal{M}_m) \prod_{j=1}^{p} \delta_0(\beta_j)^{1-\gamma_j} \tag{2}$$

(here $\boldsymbol{\beta}_m$ corresponds to the elements of $\boldsymbol{\beta}$ where $\boldsymbol{\gamma}$ equals 1). The last stage of the hierarchical model assigns prior weights to each of the models,

$$\mathcal{M}_m \sim \pi(\mathcal{M}_m) \tag{3}$$

which are often taken to be uniform *a priori*. This is a widely used non-informative choice as evidence for comparing two models using posterior model probabilities depends only on the marginal likelihood of the data, and not the prior weights assigned to the models.

By collapsing the last two stages, the marginal prior distribution for $\boldsymbol{\beta}$ is a mixture of point mass and continuous distributions defined on $(\{0\} \cup (-\infty, \infty))^p$. Likewise, the posterior distribution is also a mixture distribution, reflecting model uncertainty.

3.2 Posterior Distributions

Using Bayes Theorem, the posterior probability of model $\mathcal{M}_m$ is

$$\pi(\mathcal{M}_m|\mathbf{Y}) = \frac{f(\mathbf{Y}|\mathcal{M}_m)\pi(\mathcal{M}_m)}{\sum_{k=1}^{M} f(\mathbf{Y}|\mathcal{M}_k)\pi(\mathcal{M}_k)} \tag{4}$$

where the marginal distribution of the data is

$$f(\mathbf{Y}|\mathcal{M}_m) = \int f(\mathbf{Y}|\boldsymbol{\beta}, \mathcal{M}_m)p(\boldsymbol{\beta}|\mathcal{M}_m)d\boldsymbol{\beta} \tag{5}$$

given model $\mathcal{M}_m$, which provides a measure of how much the data support each model. Even if uniform prior model probabilities are used, the prior on $\boldsymbol{\beta}$ does impact posterior model probabilities through the marginal likelihood of the data, $f(\mathbf{Y}|\mathcal{M}_m)$.

Under BMA, the distribution of quantities of interest, Δ, such as future mortality or relative risks, is represented as a mixture distribution,

$$f(\Delta|\mathbf{Y}) = \sum_{m=1}^{M} f(\Delta_m|\mathbf{Y}, \mathcal{M}_m)\pi(\mathcal{M}_m|\mathbf{Y}) \tag{6}$$

where the model specific distributions are weighted by the posterior model probabilities.

3.3 Prior Distributions for BMA

A key component to model averaging is the prior distribution for the parameters and models; how should we choose $p(\boldsymbol{\beta}_m|\mathcal{M}_m)$ and $\pi(\mathcal{M}_m)$? Subjective prior distributions may be difficult to elicit in large problems, especially when there are complicated interactions and dependencies among variables and robustness of the prior is a concern. In complicated models, aspects of the prior distribution that have not received careful attention may lead to undesirable behavior in the posterior distribution (Berger 1985). In linear regression, $p(\boldsymbol{\beta}_m|\mathcal{M}_m)$ is often based on a normal distribution with mean zero, and prior covariance $\tau(\mathbf{X}'_m\mathbf{X}_m)^{-1}$ and $\pi(\mathcal{M}_m)$ is a uniform distribution over $\mathcal{S}$, so that all models are equally likely *a priori*. Similar priors are also often used in generalized linear models (Raftery 1996). In linear regression, this form is often selected out of convenience because the calculations under conjugate distributions lead to closed form solutions for the posterior distributions. The hyperparameter τ can have a strong impact on model selection, which is often not taken into consideration when developing a "non-informative", but proper prior. In fact, proper, but vague "non-informative" priors obtained by taking τ too large, can lead to Bayes factors favoring the null model, even in situations where the parameter estimates may be far from 0 in a practical sense (Kass and Raftery 1995).

While a careful subjective Bayesian analysis is ideal, we argue that even when subjective information is available, it may be desirable to present results based on objective prior distributions to accompany subjective analyses to check sensitivity of results to prior specifications. The Schwartz criterion or BIC has been used to provide a reference procedure for scientific reporting, and can be normalized to provide weights for BMA (Kass and Raftery 1995). AIC is another commonly used default procedure for model selection, but has not been used in conjunction with BMA. In the next section, we describe objective prior distributions for parameters and models based on a modification of Jeffreys' prior. These distributions have one hyperparameter c that can be used to calibrate the priors based on standard model selection criteria, such as AIC, BIC, and RIC. This provides a range of objective prior and posterior distributions for use in BMA and scientific reporting, requiring varying degrees of evidence.

4 Objective Prior Distributions

In estimation problems with vector valued parameters, Jeffreys (1961) suggested taking

$$p(\boldsymbol{\beta}_m | \mathcal{M}_m) = |\mathcal{I}(\boldsymbol{\beta}_m)|^{1/2} \tag{7}$$

where $|\mathcal{I}(\boldsymbol{\beta}_m)|$ is the determinant of the expected Fisher information matrix. In many problems, this leads to an improper distribution, which is determined only up to a multiplicative constant. We can always multiply an improper prior as in (7) by a constant a_m and still have a "valid" improper prior distribution. The constants do not affect the posterior distribution of $\boldsymbol{\beta}$ given $\mathcal{M}_m$, but are present in the marginal likelihood, and thus Bayes factors or posterior model probabilities contain the "arbitrary" constants.

Jeffrey's prior with specific values of these constants can be tuned or calibrated, however, so that posterior model probabilities reflect utilities under different model selection rules. Starting with Jeffreys' prior we develop below a family of Calibrated Information Criterion

(CIC) prior distributions that have a single hyperparameter c. Through specific choices of c, the CIC prior distributions can be used to reconcile classical model selection and Bayesian model selection based on posterior model probabilities. The CIC priors provide a general framework for model averaging, as opposed to model selection, when model uncertainty is an issue.

4.1 Calibrated Information Criterion Prior Distributions

For generalized linear models, we define the CIC prior distribution by modifying Jeffreys' prior in the following way:

$$p(\boldsymbol{\beta}|\mathcal{M}_m)\pi(\mathcal{M}_m) = (2\pi)^{-d_m/2} \left|\frac{1}{c}\mathcal{I}(\hat{\boldsymbol{\beta}}_m)\right|^{1/2} \prod_{j=1}^{p} \delta_0(\beta_j)^{1-\gamma_j} \tag{8}$$

where d_m is the dimension of model $\mathcal{M}_m$ and $\mathcal{I}(\hat{\boldsymbol{\beta}}_m)$ is the observed Fisher information for $\mathcal{M}_m$ evaluated at the MLE's $\hat{\boldsymbol{\beta}}_m$ with j,k elements,

$$[\mathcal{I}(\boldsymbol{\beta}_m)]_{jk} = -\left[\frac{\partial^2}{\partial\beta_j\partial\beta_k}\mathcal{L}(\boldsymbol{\beta}|\mathcal{M}_m)\right]$$

and $\mathcal{L}(\boldsymbol{\beta}|\mathcal{M}_m)$ denotes the log likelihood under $\mathcal{M}_m$. The hyperparameter c will be used in calibration of posterior model probabilities. The term involving (2π) appears in order to simplify expressions in the posterior and calibration, but otherwise could be absorbed into c.

The use of Jeffreys' prior or the determinant of the Fisher information ensures that the prior distribution automatically takes into account the link function, so that the priors can be used with the canonical log link or the identity link. The latter may be preferable when spatial variation of covariates is an issue and data have been spatially aggregated (Clyde 1999). For the Poisson regression model with the log link the observed Fisher information is

$$\mathcal{I}(\hat{\boldsymbol{\beta}}_m) = \mathbf{X}_m' V(\hat{\boldsymbol{\beta}}_m)\mathbf{X}_m \tag{9}$$

where $V(\boldsymbol{\beta}_m)$ is the covariance matrix for $\mathbf{Y}$ with elements $\exp(\mathbf{X}_m\boldsymbol{\beta}_m)$ on the diagonal and 0 elsewhere. For the canonical link, the observed and expected Fisher information evaluated at the MLE are the same. While one could use the expected information or a non-data dependent version of the prior by not replacing $\boldsymbol{\beta}_m$ by its MLE, this would lead to additional terms in the posterior model probabilities and require additional approximations in order to obtain the desired calibration of posterior model probabilities and model selection criteria.

4.2 CIC Posterior Distributions

For the Poisson regression models under consideration we cannot obtain the marginal likelihood of the data (5) analytically. Laplace's method (Tierney and Kadane 1986) provides a useful approximation to the marginal likelihood as long as the likelihood is peaked near its maximum, which will be the case for large samples. Kass and Raftery (1995) have found that Laplace approximations for determining posterior model probabilities are accurate for sample sizes on the order of $20p$ or larger. As in classical approximations for obtaining the distribution of MLE's, we replace $\mathcal{L}(\boldsymbol{\beta}|\mathcal{M}_m)$ by a 2nd order Taylor series expansion about $\hat{\boldsymbol{\beta}}$, so that under $\mathcal{M}_m$

$$\tilde{\mathcal{L}}(\boldsymbol{\beta}|\mathcal{M}_m) = \mathcal{L}(\hat{\boldsymbol{\beta}}|\mathcal{M}_m) - \frac{1}{2}(\boldsymbol{\beta}_m - \hat{\boldsymbol{\beta}}_m)'\mathcal{I}(\hat{\boldsymbol{\beta}}_m)(\boldsymbol{\beta}_m - \hat{\boldsymbol{\beta}}_m).$$

Because the prior is independent of β, this is equivalent to a Laplace expansion of the posterior distribution around the posterior mode. Using $\exp(\tilde{\mathcal{L}}(\boldsymbol{\beta}|\mathcal{M}_m))$ in place of the actual likelihood, the (approximate) joint posterior for $\boldsymbol{\beta}_m$ and $\mathcal{M}_m$ factors as

$$p(\boldsymbol{\beta}_m|\mathbf{Y}, \mathcal{M}_m) = N(\hat{\boldsymbol{\beta}}_m, \mathcal{I}(\hat{\boldsymbol{\beta}}_m)^{-1}) \tag{10}$$

$$\pi(\mathcal{M}_m|\mathbf{Y}) = \frac{\exp\left\{\frac{1}{2}\left(D_m - d_m\log(c)\right)\right\}}{\sum_{m=1}^{M}\exp\left\{\frac{1}{2}\left(D_m - d_m\log(c)\right)\right\}} \tag{11}$$

where D_m is the model deviance, which is the usual deviance (-2 times the log likelihood) under the null model minus the deviance under $\mathcal{M}_m$.

The log of the posterior model probability under the CIC prior distribution is

$$\log \pi(\mathcal{M}_m|\mathbf{Y}) = k + \frac{1}{2}(D_m - d_m\log(c))$$

where k is the normalizing constant. The log posterior model probability is proportional to the model deviance minus a model complexity penalty term depending on $\log(c)$. Posterior model probabilities under the CIC priors can be calibrated to classical model selection through the choice of $\log(c)$ for popular methods such as AIC (Akaike 1973, 1978), BIC (Schwarz 1978) and RIC (Foster and George 1994) (Table 2). For the values of $\log(c)$ in Table 2, the model with the highest posterior probability under that prior corresponds to the optimal CIC model using a penalized deviance criterion for model selection. Using the CIC prior distributions, posterior inference within a model is based on standard normal approximations in GLMs given by (10), but the CIC posterior model probabilities can also be used to incorporate model uncertainty using (6).

[Table 2 here]

While much of statistical practice has focused on selecting a single model, this may be unreasonable unless $\pi(\mathcal{M}_k|\mathbf{Y})$ is near one for one of the models under consideration or the best models provide similar inferences; even if the posterior probability of the best model is near one, there is little loss by using BMA as the best model dominates the mixture. Raftery (1996) used BMA based on BIC in GLMs, however, the usual justification of AIC provides no way of taking into account model uncertainty (Kass and Raftery 1995). The CIC prior distributions do provide a justification for model averaging using AIC and other classical model selection criteria based on penalized deviance criteria of the form above.

Our goal in using the CIC prior formulation is to provide a family of prior distributions for sensitivity analyses with BMA. At one extreme, AIC is a popular model selection procedure (it is the default in SPlus) and often favors complex models, while at the other extreme, BIC is a more conservative strategy and requires much stronger evidence to accept the alternative hypothesis. Both approaches are not without critics. AIC often tends to overestimate the number of parameters needed, even asymptotically (see discussion in Kass and Raftery 1995). This, however, depends critically on the asymptotics of how the design matrix changes with sample size. For finite dimensional models, BIC is known to be asymptotically consistent and prediction optimal, while AIC is not; for infinite dimensional models AIC gives optimal prediction while BIC does not (Hansen and Yu 1999). There is empirical evidence that BIC may be too conservative, leading to elimination of variables that have real, but small effects (see Weakliem 1999 and discussion). In a one dimensional testing problem, t statistics in favor of the alternative under AIC correspond to t^2 values greater than 2 while for BIC, t^2 values must be greater than $\log(n)$; RIC is often in between, where t^2 values must exceed

$2\log(p)$. Because of the multiple testing framework of model selection, the overall Type I error rate of AIC is much higher than the rate for a single test. RIC compensates by adjusting for the number of parameters (Foster and George 1994). For finite samples, Fernández *et al.* (1998) conducted simulation studies in linear models under several model, prior and sample size scenarios, and found that BMA based on RIC priors provided better performance for sample sizes less than p^2, while BIC was better with larger sample sizes. Simulation studies, however, cannot cover all contingencies. While the CIC prior distributions may be calibrated to model selection or predictive criteria such as AIC, BIC, and RIC, other approaches for specifying the constants can be based on imaginary training as in Spiegelhalter and Smith (1982). This may be useful for judging whether BIC leads to results that are too conservative. In general, there is no consensus regarding the best "default" prior for model selection and model averaging. However, by using a range of values for c we can judge how sensitive inferences under BMA are to prior specifications.

A high probability of variable inclusion under both AIC and BIC priors implies consistent strong support for an effect, while if both AIC and BIC lead to small probabilities of variable inclusion, this provides consistent support in favor of the null. Situations where the probability of variable inclusion is high under AIC, but low under BIC, may indicate that the sample is not large enough to detect the effect, and that conclusions will be sensitive to the choice of c or choice of other subjective prior distributions.

4.3 Implementing Model Averaging

For linear regression models with conjugate prior distributions and a small to moderate number of covariates (less than 20), posterior distributions for many quantities can be determined analytically (George and McCulloch 1997). For larger problems, we typically cannot enumerate all models so model averaging is approximated by using a sample of models from $\mathcal{S}$. Stochastic search using Markov Chain Monte Carlo (MCMC) methods or deterministic search methods such as leaps and bounds can be used to identify a sample of models that are used in BMA (see George and McCulloch 1997, Clyde 1999 and Hoeting *et al.* 1999 for discussion of approaches and methods for implementing BMA in the context of linear and generalized linear models). For large problems with highly correlated variables (such as the meteorological variables) using transformations based on factor analyses or principal components may lead to improved convergence with MCMC methods (Clyde 1999, Clyde and DeSimone-Sasinowska 1997).

For the application in the next section, we modified the `bic.glm` code (available on the BMA homepage, URL `http://att.research.com/~volinsky/bma.html`), written by C. Volinsky. This uses the leaps and bounds algorithm to provide a preliminary list of models for use with the objective CIC prior distributions; if a more detailed analysis including subjective information or other proper priors is warranted, then one may later implement Monte Carlo or MCMC methods to provide posterior samples for making inferences.

5 Results for Birmingham

We explore the use of the CIC prior distributions for assessing the effect of particulate matter on elderly mortality in Birmingham, AL and what impact model uncertainty may have on decisions. We use the Poisson model with canonical log link given in (1) and consider possible models based on the variables listed in Table 1.

As the combined design matrix for the thin-plate spline basis and meteorological and PM_{10} variables contains almost 60 variables, and the leaps and bounds code in SPlus is limited to 30 variables, one must either eliminate variables or use a multistage procedure. We implemented the model search in two stages: the first stage was used to estimate the non-parametric baseline trend using thin-plate smoothing splines; the second stage included the posterior mean of the baseline trend estimated in Stage 1 and all meteorological and PM_{10} variables listed in Table 1.

5.1 Baseline Trend Estimates

Figure 3 illustrates the degree of model uncertainty in the estimates of the baseline trend. The thin solid line is the GLM estimate under the full model with all 30 knots. This closely tracks a number of the high mortality episodes which may be a result of other (short-term) factors. The dashed lines correspond to the top 100 models under the CIC posterior using $c = n$, with the thick solid line corresponding to the predictive mean under BMA. This provides an objective estimate of the baseline trend, without subjective assessment of the number of knots, and appears to capture the necessary long-term variation without over-fitting. We incorporate the BMA estimate of the baseline trend as a linear predictor in the model; this can be thought of as an independent underlying baseline estimate as in proportional hazards models. While this two stage approach ignores uncertainty in the baseline estimate, we can later account for the additional uncertainty by repeating Stage 2 for a number of the top models from Stage 1, or by re-fitting models identified in Stage 2 under the baseline models identified in Stage 1.

[Figure 3 here]

5.2 Covariate and Lag Choice

In Stage 2, 7860 models were selected by the leaps and bounds approach for use in BMA. We used all 860 of the observations that had complete records for all of the lagged `pm` variables, so that with all lags and the BMA baseline estimate the design matrix for the full model included 29 candidate predictors. The variable selection penalties $\log(c)$ under RIC and BIC are very close, 6.73 versus 6.75 respectively, so that results are reported using BIC only.

Relative risks for each model were based on a simultaneous 100 $\mu g/m^3$ increase in all PM_{10} variables (`pm0, pm1, pm2, pm3`)

$$R = \exp\left(100(\beta_{\texttt{pm0}} + \beta_{\texttt{pm1}} + \beta_{\texttt{pm2}} + \beta_{\texttt{pm3}})\right) \tag{12}$$

so that results could be compared to Schwartz (1993). The posterior distribution for the relative risk given $\mathcal{M}_m$ is a log-normal distribution

$$\log(R)|\mathcal{M}_m, \mathbf{Y} \sim N\left(100(\beta_{\texttt{pm0}} + \beta_{\texttt{pm1}} + \beta_{\texttt{pm2}} + \beta_{\texttt{pm3}}), \sigma^2_R\right)$$

with variance

$$\sigma^2_R = 100^2\left(1'\mathbf{\Sigma}_{\boldsymbol{\beta}_{\text{pm}}|\mathcal{M}_m}1\right)$$

where $1' = (1,1,1,1)'$ and $\mathbf{\Sigma}_{\boldsymbol{\beta}_{\text{pm}}|\mathcal{M}_m}$ is the covariance matrix for the PM_{10} regression coefficients under model $\mathcal{M}_m$ derived from the Fisher information (9) under model $\mathcal{M}_m$ with zero entries for the degenerate components for lags not included under $\mathcal{M}_m$. Under the normal model, Bayesian probability intervals conditional on a model correspond to classical confidence intervals used in other analyses.

5.3 Model Uncertainty

Figure 4 “links” plots of the model space and corresponding MLE's of relative risks (posterior modes) under the top 25 models under the BIC and AIC priors. The model space is represented as a matrix, where rows correspond to models and columns to variables. In the model matrix, a black square in position jk indicates that the kth variable is not included in the jth model. The models are ordered from best (at the top) to worst (on the bottom) with the scale on the y-axis reflecting the log ratio of the model probability of the best model to the worst model in the entire sample of models. A difference of 2 or less in this scale indicates that the top 25 models are more or less exchangeable.

[Figure 4 here]

The last 4 columns in the model space matrix correspond to the PM_{10} variables: `pm0 - pm3`, the current day and 3 previous days. As the figure shows, the top models under the AIC and BIC priors are very different regarding the inclusion of PM_{10}: PM_{10} variables are included in all of the top AIC models, but in roughly one third of the top 25 BIC models. The y-axis in the relative risk plots in Figure 4 also corresponds to models with the best model at the top, thus providing a static link between the model space, which identifies the confounding variables, and the corresponding relative risk estimates. Clearly models where all the coefficients for PM_{10} are zero have relative risks of 1. The points indicate the mean relative risk under each model, while the horizontal lines correspond to 95% probability (confidence) intervals. While the top 25 AIC models include PM_{10}, most of the probability (confidence) intervals for relative risk include one. Under both AIC and BIC there are two models in the top 25 that have probability intervals for relative risk that exclude 1. However, without additional scientific (subjective) input, there is no reason based on the support they receive from the data to prefer these over any of the other top models.

5.4 Distribution of Relative Risk

Figure 5 illustrates the conditional distribution of relative risk over all sampled models given that PM_{10} or some lag of PM_{10} was included in the model under both the BIC and AIC priors. The histograms at the top of the figure show the distributions of the MLE's of relative risks weighted by conditional posterior model probabilities (given that the relative risk is not one). This illustrates model uncertainty in the (point) estimates for models with PM_{10}, but ignores parameter uncertainty. The range in the estimates (indicated by the dots) is almost as large as the length of individual probability intervals in Figure 4. The histograms at the bottom of Figure 5 are based on samples of relative risks from the posterior distribution under model averaging over models with PM_{10}, which incorporates both model uncertainty and parameter uncertainty. While the range and values of the relative risks are the same under both the AIC and BIC posteriors (the same sampled models and realizations are used to reduce Monte Carlo variation), the weights used in constructing the histograms depend on the posterior model probabilities and priors. The posterior model probabilities under AIC tend to be more dispersed, leading to increased model uncertainty; this in turn is reflected in more uncertainty in the distributions of the relative risk.

The unconditional distribution of the relative risk (over all models) includes a "spike" at relative risks of one corresponding to models with no PM_{10} variables; the height of the spike depends on the posterior probability that the relative risk is one (Table 5). Under AIC, the posterior probability that the relative risk is 1 given the data is 0.03, while under the BIC prior the posterior probability that the relative risk is 1 given the data is 0.72. Because of the difference in the mass at relative risks for one, the unconditional relative risk distributions under the AIC and BIC posterior distributions are quite different.

[Figure 5 here]

Table 3 summarizes relative risks using the AIC and BIC prior distributions under model averaging. While posterior means for the relative risk (given that PM_{10} or some lag of PM_{10} is included in the model) are comparable under the BIC and AIC priors, (both approximately 1.05), the uncertainty over whether PM_{10} variables should be included in the model depends greatly on whether the AIC or BIC priors are used, and in turn impacts the overall estimate of the relative risk averaged over all sampled models (1.052 for AIC versus 1.015 for BIC). The 95% probability interval from the mixture distribution from BMA under the BIC prior is (0.99, 1.11) while for the AIC prior the interval is (0.95, 1.18). While the overall mean relative risk is higher under BMA with the AIC prior, the associated level of uncertainty is also greater, resulting in a wider probability interval. These are based on a 100 unit change in PM_{10} for direct comparison with Schwartz, who obtained a relative risk of 1.11. While the results of Schwartz are plausible (there are models in the top 25 that have relative risks in this range), results under both the AIC and BIC priors appear to support lower estimates of relative risks when model uncertainty in taken into account. A 100 $\mu g/m^3$ change in PM_{10} levels is unrealistic as the range of PM_{10} in Birmingham over the study period was 8 to 163 $\mu g/m^3$. Relative risks based on a 10 $\mu g/m^3$ change are more realistic, leading to intervals of (0.995, 1.016) under AIC and (0.999, 1.011) under BIC using BMA. Similarly, relative risks based on a one interquartile range change (36 $\mu g/m^3$) are (0.981 ,1.061) and (0.996, 1.039) under the AIC and BIC posterior distributions, respectively.

[Table 3 here]

We also estimated relative risks using the area-wide average of PM_{10} and lags, and found that the estimates of relative risk under BMA were 1.02 for AIC and 1.009 for BIC. SDSSS have looked at area wide exposure measures for daily PM_{10} that exclude the Leeds data, and have also found that relative risk estimates are higher when those monitors are excluded, although the differences between their results under the area wide exposure measure with and without the Leeds monitor are not great.

5.5 Posterior Effect Probabilities

We may fail to reject the null hypothesis of "no effect" for PM_{10} because either there is not enough data to detect an effect, or the data actually support the null hypothesis; one advantage of BMA over the use of traditional p-values is that posterior probabilities can distinguish between these two situations (Hoeting *et al.* 1999). The probability of no effect is equivalent to the posterior probability that the relative risk is 1. Under the BIC prior, the posterior probability is 0.72, so that the data are weakly in favor of no effect, while under the AIC prior, the probability is 0.03, strong evidence against a relative risk of 1. While the results are sensitive to the prior distribution, in both cases, the data do not provide strong convincing evidence that there is no PM_{10} effect. While the AIC prior results in more support for the hypothesis that the relative risk is greater than 1 (posterior probability = 0.967) compared to BIC (posterior probability = 0.258), both approaches indicate that the increase in mortality is likely much less than 11 percent; the posterior probability that the relative risk is greater than 1.11 is 0.042 and 0.007 under AIC and BIC respectively.

5.6 Validation Study

As the importance of the effect of PM_{10} on mortality depends on the choice of prior, we carried out a small validation study to compare the predictive performance of Bayesian model averaging and model selection under the AIC and BIC prior distributions. For this we randomly selected 75 days with complete PM_{10} data, and repeated the BMA analysis described in the previous section using the remaining data. For the 75 days in the validation set, we computed the predictive MSE,

$$\text{MSE} = \sum_{i=1}^{75} (Y_i - \hat{Y}_i)^2/75$$

where $\hat{Y}_i$ is the predictive mean of daily elderly non-accidental mortality. We calculated predictive means and MSE's under BMA with the AIC and BIC priors, and for the best AIC and BIC models (Table 4). In both cases, model averaging leads to better predictive performance than model selection. The predictive comparison favors model averaging under the BIC priors, with a gain in efficiency over the best AIC model of just over 5%. Similar results were obtained with other randomly selected validation sets. BIC acts like a fully automatic Occam's razor by cutting back to the simpler model when little is lost for predictive purposes (Kass and Raftery 1995); this is, however, sometimes at the expense of rejecting the true model. Because the relative risk estimates are small, there may be little lost for predictive purposes by using the simpler model without PM_{10} (the best model under BIC). However, BMA with BIC provides better predictive performance than model selection under BIC, supporting evidence that there is a small effect. While this comparison supports the simpler models under BMA with BIC over AIC, little research has been conducted on "validating" PM_{10} mortality health effect models, and this is an open area for design of appropriate methods.

[Table 4 here]

6 Discussion

With the potential for reanalysis of existing studies due to legal challenges of the new standards, model averaging provides a coherent methodology for combining inferences from different models for the same data, as well as judging how well the data support competing models. The use of Bayesian model averaging can help eliminate the concern that observed "associations" between increased particulate matter and mortality are an artifact of multiple selection. By explicitly considering model uncertainty in analyses, more realistic measures of uncertainty for relative risks can be obtained, providing a more secure foundation for decision making.

Using the methodology presented here, BMA can become a routine part of exploratory data analysis as most quantities of interest are based on standard output from GLM packages. The distributions of relative risk under BMA require generation of relative risks from the posterior distribution, but this is straightforward under most higher level statistical packages. The results are all conditional on the collection of models used, which does require special consideration. We hope that by exploring linked plots we can gain a better understanding of confounding variables and given the collection of models, one can objectively discuss the relative merits of the top models. As we proceed, we may find it necessary to enlarge the class of models under consideration by allowing for overdispersion, nonlinear effects, interactions, measurement errors, latent variables, and spatial variation. Bayesian model averaging can be applied sequentially as new information or variables become available, and can be used to take into account other structural features of the model beyond covariate choice.

In this example, it is clear that the relative risk estimates vary among the top models, and that the importance of the PM_{10} effect depends on the prior distribution that is used. Our goal in using the CIC formulation for BMA is primarily for sensitivity analyses for comparison with other classical based analyses or Bayesian analyses, and thus we suggest reporting estimates under a range of priors encompassing AIC, RIC, and BIC. The sensitivity analyses using the CIC priors provide some evidence of whether the conclusions depend on subjective prior assumptions or model selection strategies. If results using CIC and subjective priors are consistent, even though a subjectively based prior may not be acceptable to the majority of researchers or consumers of PM_{10} research, we are more confident in the conclusions. On the other hand, if BIC does not provide evidence in favor of an effect, while a subjective or classical approach does, then we need to look carefully at the assumptions and justification of the prior distributions, but also be aware that BIC is sometimes too conservative.

In the situation where a decision maker does not want to incorporate subjective information (or it is not available), and AIC, RIC, and BIC lead to different conclusions, Empirical Bayes (EB) prior distributions for model averaging are an appealing addition to the CIC prior distributions (Clyde and George 1998, George and Foster 1997). The EB prior distributions allow the data to adaptively determine the hyperparameters, and ensure that the prior distribution is not at odds with the data. Related to this is the Minimum Description Length (MDL) approach to model selection of Hansen and Yu (1999) that uses data-dependent adaptive penalties for model selection. These adaptive approaches have provided excellent predictive performance in linear models in comparison to AIC, BIC and RIC, and can provide values of c that are consistent with the observed data. Future work includes the development of EB prior distributions for Poisson models and other generalized linear models.

We have considered how model uncertainty may impact conclusions within the the framework of epidemiological models currently under consideration in practice, with a focus on covariate and lag choice. Because the analyses are based on observational data, there are a number of reasons why we may observe associations between health outcomes and particulate matter levels, besides a direct causal relationship. Missing confounders and other latent variables provide other explanations for associations that cannot be ruled out. Because we do not directly observe an individual's exposure and the area-wide averages are a proxy for exposure, latent variable processes for exposure that incorporate various sources of measurement error may provide more accurate estimates of relative risks combined with using an identity link to preserve estimates under spatial aggregation. Structural equation models may provide a refinement to the current class of epidemiological models that allow one to investigate causal theories in conjunction with latent variables. While Bayesian model averaging can theoretically be used to take into account uncertainty in latent and direct causal structures, claims of causality based on observational studies may be highly sensitive to the choice of prior distributions and class of models under consideration. While causal interpretations of these models are limited because of the observational nature of the data, exploration and additional research in this area may provide new and interesting insights.

References

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood". In *2nd International Symposium on Information Theory*, Budapest: Akademia Kaido. 267-281.

Akaike, H. (1978). "A new look at the Bayes procedure". *Biometrika* 65, 53–9.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer, New York.

Clyde, M. (1999). "Bayesian model averaging and model search strategies". In *Bayesian Statistics 6* J.M. Bernardo, A.P. Dawid, J.O. Berger, and A.F.M. Smith eds. Oxford University Press. pages 157-185.

Clyde, M. and DeSimone-Sasinowska, H. 1997. "Accounting for model uncertainty in Poisson regression models: Does particulate matter particularly matter?" ISDS Discussion Paper 97-06.

Clyde, M. and George, E. (1998). "Flexible empirical Bayes estimation for wavelets". Institute of Statistics and Decision Sciences, Duke University, Discussion Paper 98–21. *http://www.isds.duke.edu*

Draper, D. (1995). "Assessment and propagation of model uncertainty (with Discussion)". *Journal of the Royal Statistical Society, Series B*, 56, 45-98..

Fernández, C., Ley, E., and Steel, M. F. (1998). Benchmark priors for Bayesian model averaging. Technical report, Department of Econometrics, Tilburg University, the Netherlands.

Foster, D.P. and George, E.I (1994). "The risk inflation criterion for multiple regression". *Annals of Statistics*, **22**, 1947–75.

George, E.I. and Foster, D.P. (1997). Calibration and empirical Bayes variable selection. Technical Report, Dept. of MSIS, University of Texas at Austin.

George, E.I. and McCulloch, R. (1997). "Approaches for Bayesian variable selection". *Statistica Sinica* **7**, 339-374.

Hanson, M. and Yu, B. (1999). Model Selection and the Principle of Minimum Description. Bell Labs Technical Report, URL *http://cm.bell-labs.com/who/cocteau/papers*

Hodges, J.S. (1987)." Uncertainty, policy analysis, and statistics". *Statistical Science* 2, 259–291.

Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). "Bayesian model averaging: A tutorial". To appear in *Statistical Science.*

Holmes, C.C. and Mallick, B.K. (1997). "Bayesian radial basis functions of unknown dimension". Tech Report, Imperial College.

Jeffreys, H. (1961). *Theory of Probability* 3rd Edition, Oxford University Press.

Kass, R.E. and Raftery, A.E. (1995). "Bayes factors". *Journal of the American Statistical Association*, 90, 773-795.

Lamon, E.C. and Clyde, M.(1998). "Accounting for model uncertainty in prediction of chlorophyll *a* in Lake Okeechobee". ISDS Discussion Paper 98-42

National Research Council. (1998). "Research priorities for airborne particulate matter". National Academy Press. Washington, DC.

Raftery, A.E. (1996). "Approximate Bayes factors and accounting for model uncertainty in generalized linear models". *Biometrika* 83, 251-266.

Schwartz, J. (1993), "Air Pollution and daily mortality in Birmingham, Alabama". *American Journal of Epidemiology*, **137**, 1136–1147.

Schwarz, G. (1978). "Estimating the dimension of a model". *Annals of Statistics* 6, 461-464.

Smith, R., Davis, J. Sacks, J. Speckman, P. , and Styer. P. (1999), "Air pollution and daily mortality in Birmingham, Alabama: A reappraisal". Tech Report. Department of Statistics, University of North Carolina.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). " Bayes factors for linear and loglinear models with vague prior information". *Journal of the Royal Statistical Society, Series B*, 44,377–387

Tierney, L. and Kadane, J.B. (1986). "Accurate approximations for posterior moments and marginal densities". *Journal of the American Statistical Association*, 81, 82–86.

Viallefont, V. Raftery, A.E. and Richardson, S. (1998). "Variable selection and Bayesian model averaging in case-control studies". Tech Report 343, Dept. of Statistics, University of Washington.

Weakliem, D.L. (1999). "A critique of the Bayesian information criterion for model selection". *Sociological Methods and Research* 27, 359-397.

PM_{10} current day, one, two, and three day lags.

Daily Monitor 0023 `pm0, pm1, pm2, pm3`

Area Wide Average `pma0, pma1, pma2, pma3`

Temperature (current day, one and two day lags)

daily minimum temperature (`tmin, tmin1, tmin2`)

daily maximum temperature (`tmax, tmax1, tmax2`)

average daily temperature (`mntp, mntp1, mntp2`)

Humidity (current day, one and two day lags)

average dew point temperature (`dptp, dptp1, dptp2`)

daily minimum relative humidity (`mnrh, mnrh1, mnrh2`)

daily maximum relative humidity (`mxrh, mxrh1, mxrh2`)

average daily specific humidity (`mnsh, mnsh1, mnsh2`)

Atmospheric Pressure (current day, one and two day lags)

average daily station pressure (`pres, pres1, pres2`)

Seasonal Trend

thin-plate spline basis with up to 30 potential knots

Posterior mean of trend `baseline`

Table 1: Explanatory variables used in the Birmingham, AL analysis.

$\log(c_m)$	CIC	Criterion
1	R^2	maximum R^2
2	AIC	Akaike Information Criterion
$\log(n)$	BIC	Bayes Information Criterion
$2\log(p)$	RIC	Risk Inflation Criterion

Table 2: Calibrated Information Criterion priors

SUMMARIES	AIC	BIC
P(Relative Risk = 1 | data)	0.03	0.72
Posterior Mean of Relative Risk for Models with PM_{10}	1.054	1.053
Posterior Mean of Relative Risk	1.052	1.015
95% Posterior Probability Interval of Relative Risk	(0.94, 1.17)	(0.99, 1.11)
Relative Risk for Best Model	1.025	1.00

Table 3: Summaries of the distribution of relative risk associated with a 100 unit increase in PM_{10} under Bayesian model averaging (BMA) and model selection using the AIC and BIC prior distributions.

Predictive MSE	AIC	BIC
Model Selection	16.83	16.03
Bayesian Model Averaging	16.31	**15.98**

Table 4: MSE for predicting mortality for the validation set under Bayesian model averaging and model selection using the AIC and BIC prior distributions.

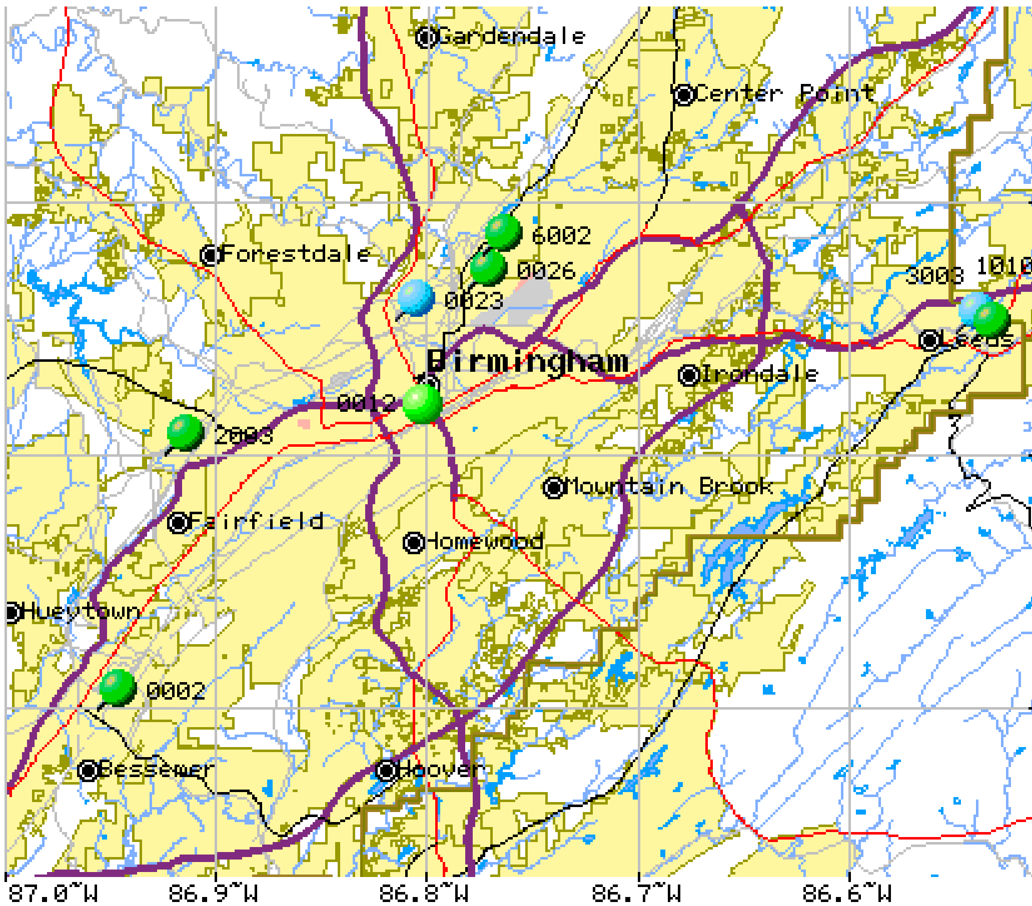

Figure 1: Location of PM_{10} monitors within the Birmingham metropolitan area (shaded area).

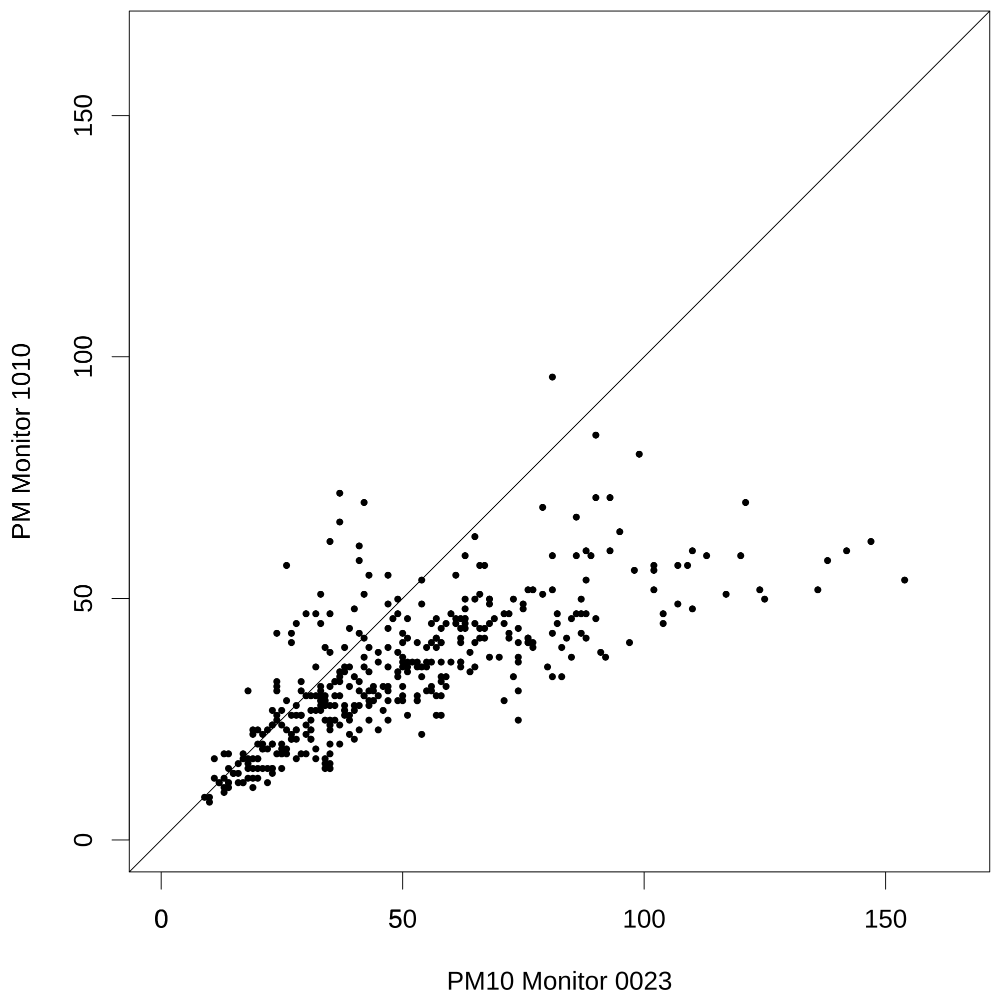

Figure 2: PM_{10} measurements for the daily monitor 1010 in Leeds versus PM_{10} measurements from the daily monitor 0023 in Birmingham.

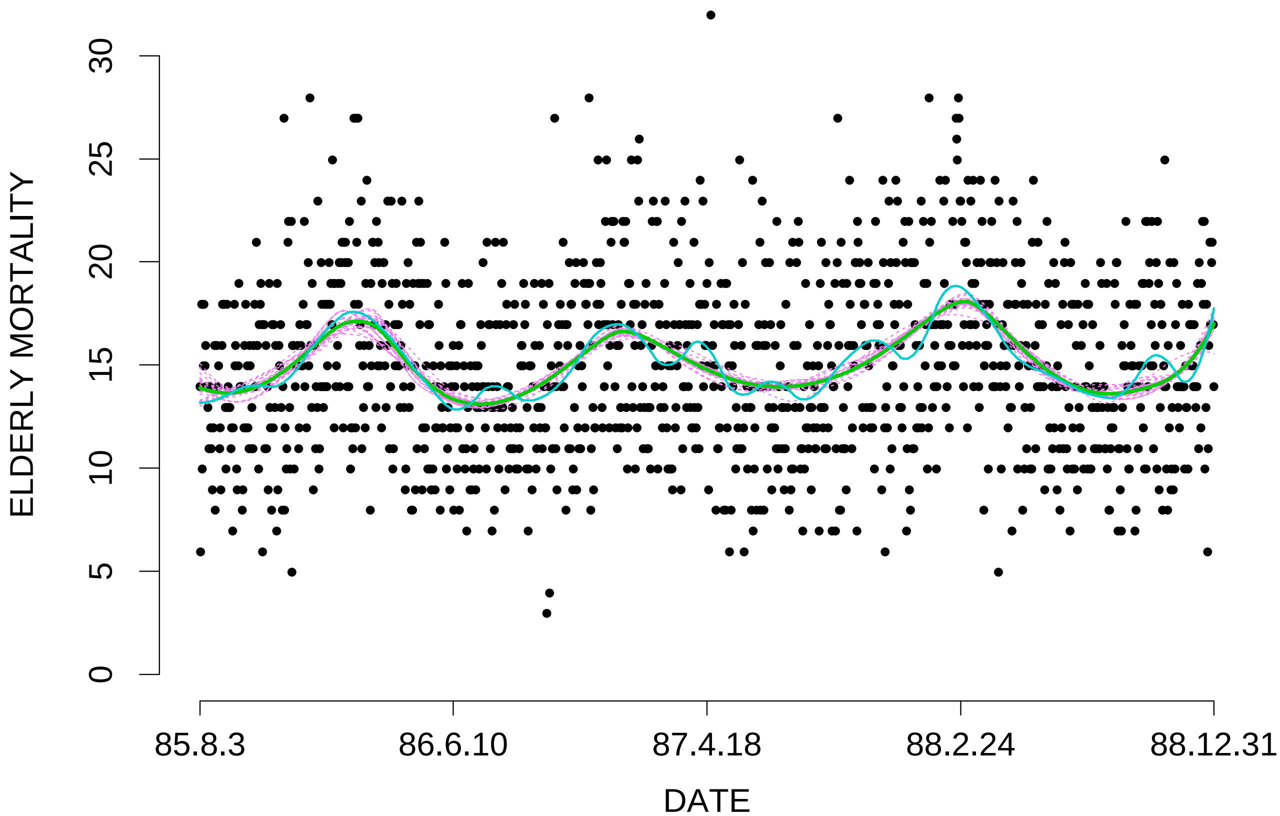

Figure 3: Plot of daily non-accidental mortality in the elderly (over 65) population for Birmingham, Alabama. The thick solid line corresponds to the baseline trend estimate under BMA with the BIC prior distribution; the thin solid line is the GLM estimate under the 30 knot model (roughly one knot for every 40 days); and the light dashed lines correspond to individual estimates from the top 100 models under the BIC prior distribution.

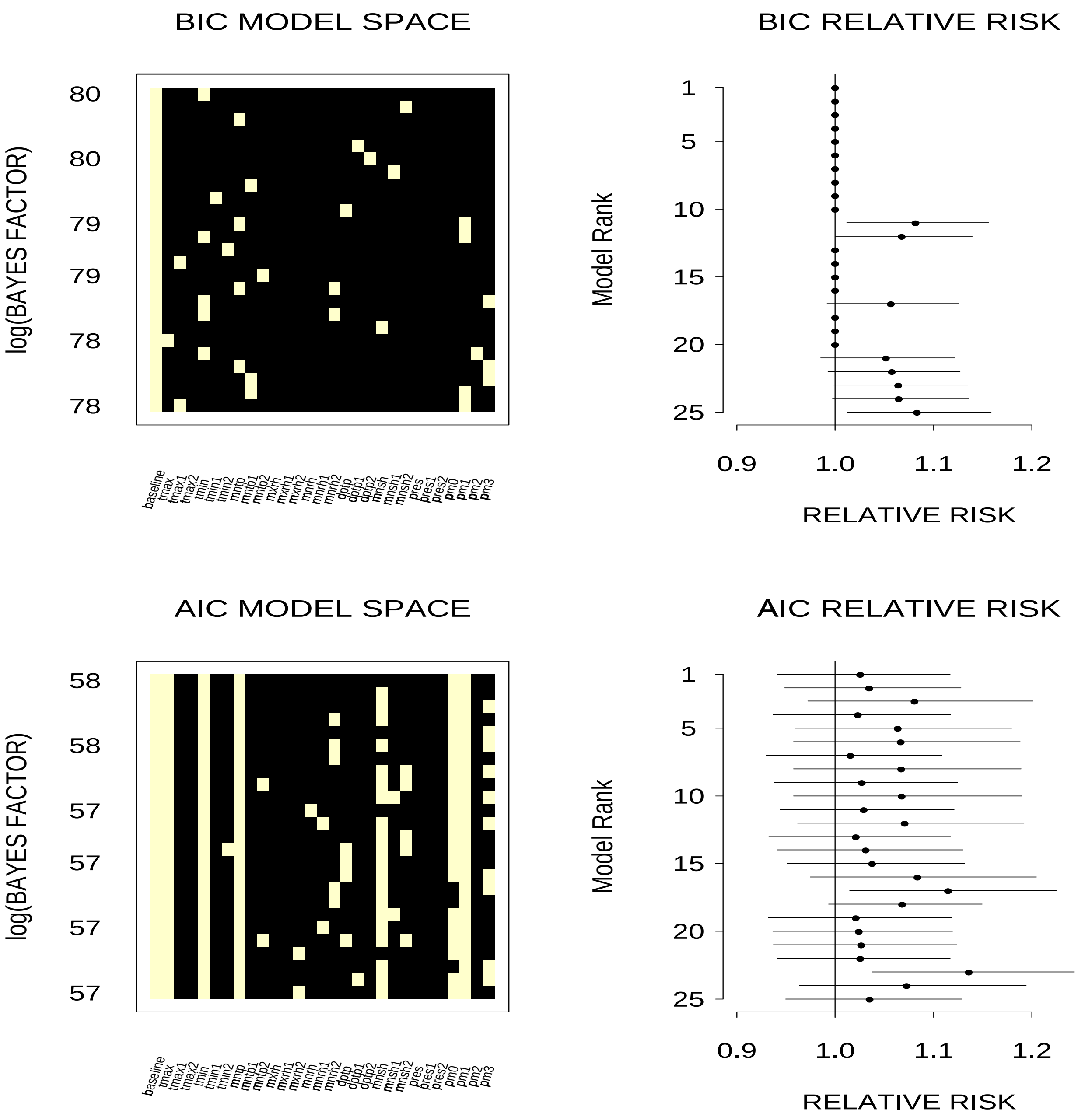

Figure 4: Top 25 models and the corresponding estimates of relative risks with 95% probability intervals under the CIC priors with $c = n$ (BIC) and $c = \exp(2)$ (AIC). Dark squares indicate that the variable in that column is not included in the model for that row.

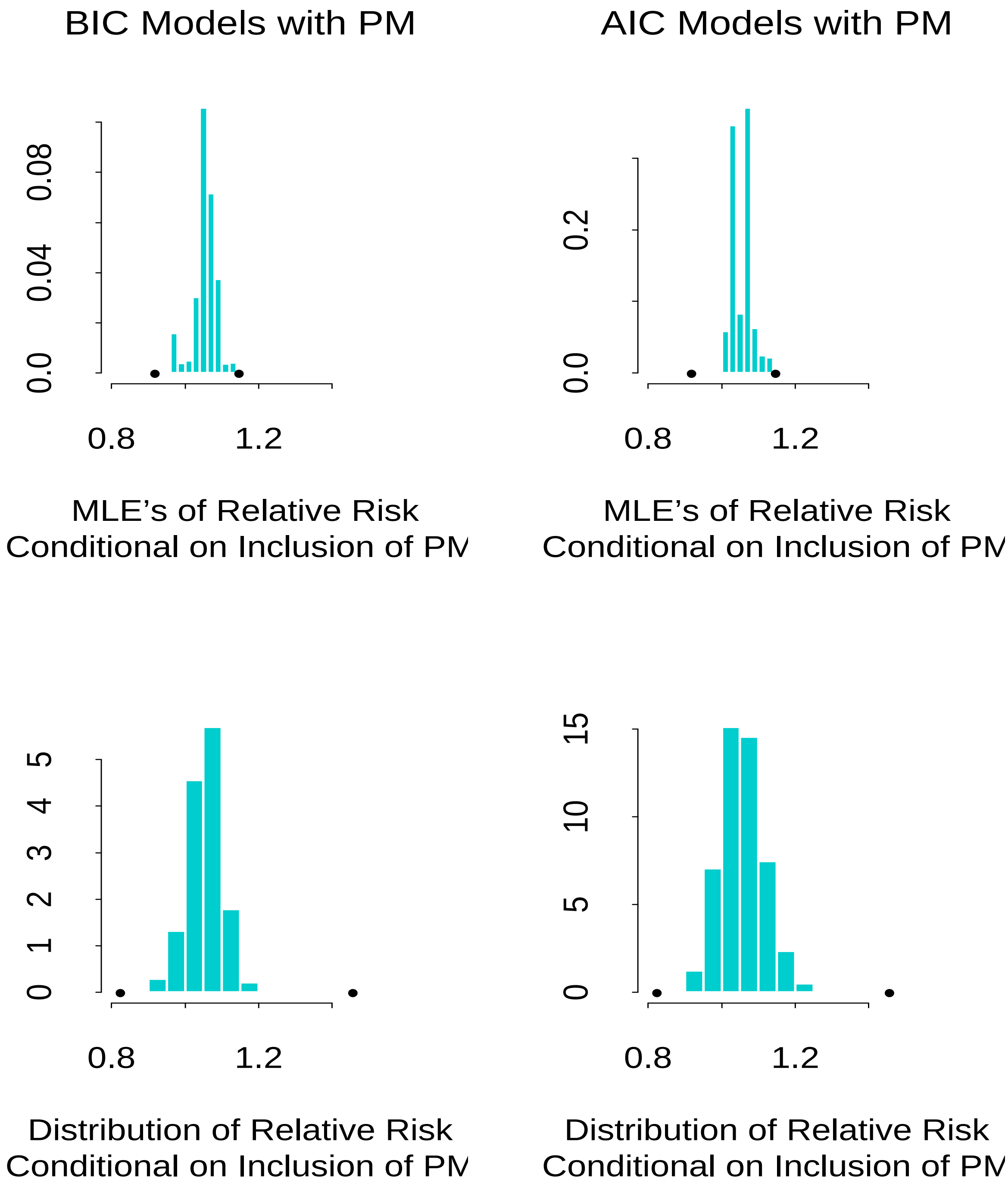

Figure 5: Conditional distribution of relative risk associated with a 100 $\mu g/m^3$ increase in PM_{10} given inclusion of PM_{10} in the models under the BIC prior (left) and AIC prior (right). The posterior probability that PM_{10} variables are included is 0.28 and 0.97 under the BIC and AIC prior distributions, respectively. Top: histograms are posterior modes (MLE) for each model weighted by respective model probabilities. Lower: histograms of relative risk incorporating both model uncertainty and parameter uncertainty. The points indicate the range of the distribution.