University of Leicester

# DEPARTMENT OF ECONOMICS

# AN INVESTIGATION OF THRESHOLDS IN AIR POLLUTION-MORTALITY EFFECTS

**Gary Koop**
**University of Leicester**

**Lise Tole**
**University of Leicester**

**Working Paper No. 04/20**
**May 2004**

# An Investigation of Thresholds in Air Pollution-Mortality Effects

Gary Koop & Lise Tole
Department of Economics
University Road, Leicester University
Leicester, U.K. LE1 7RH
lat7@le.ac.uk; phone: +44(116) 252-2503

May 2004

Abstract: In this paper we introduce and implement new techniques to investigate threshold effects in air pollution-mortality relationships. Our key interest is in measuring the dose-response relationship above and below a given threshold level where we allow for a large number of potential explanatory variables to trigger the threshold effect. This is in contrast to existing approaches that usually focus on a single threshold trigger. We allow for a myriad of threshold effects within a Bayesian statistical framework that accounts for model uncertainty (i.e. uncertainty about which threshold trigger and explanatory variables are appropriate). We apply these techniques in an empirical exercise using daily data from Toronto for 1992-1997. We investigate the existence and nature of threshold effects in the relationship between mortality and ozone ($O_3$), total particulate matter (PM) and an index of other conventionally occurring air pollutants. In general, we find the effects of our considered pollutants on mortality to be statistically indistinguishable from zero with no evidence of thresholds. The one exception is ozone, for which results present an ambiguous picture. Ozone has no significant effect on mortality when we exclude threshold effects from the analysis. Allowing for thresholds we find a positive and significant effect for this pollutant when the threshold trigger is the average change in ozone two days ago. However, this significant effect is not observed after controlling for PM.

Key Words: Threshold-air pollution mortality effects, Bayesian model averaging, PM, $O_3$.

Acknowledgements: The authors wish to thank Phil Kiely, Ontario Ministry of the Environment; Bryan Smith, Ontario Climate Centre, Environment Canada; Dr. Elizabeth Rael, Public Health Branch, Ontario Ministry of Health and Long-Term Care; and Tom Dan, Environment Canada, for their help with the data. We would also like to thank Janet Phillips, Toronto Department of Public Health, for her expert assistance with the mortality data and for her many valuable responses to our requests for information.

# 1 Introduction

Early epidemiological studies on the health effects of air pollution were concerned with measuring the link between daily deaths and the kind of severe pollution episodes that occurred in London England in 1952, Donora Pennsylvania in 1948, and the Meuse Valley Belgium in 1930. Using simple methods, these studies established a link between cardiopulmonary mortality and extreme levels of sulfur oxide and particulate matter. The alarmingly high number of deaths attributed to these pollution episodes motivated subsequent public policy efforts to regulate air pollution. As concentrations of pollutants in the air have fallen, researchers have turned their attention to measuring mortality effects at commonly encountered levels of exposure. The shape of the dose-response curve has also become a key focus of research and policy concern.[1]

In the past decade, advances in statistical techniques and computing power have enabled researchers to test these relationships more rigorously, particularly in respect to temporal confounders. Generalized linear models (GLMs) and generalized additive models (GAMs) using nonparametric smoothing techniques are now widely applied in the analysis of pollution health effects. These methods are more flexible than previous techniques in their treatment of long term trends, seasonality and other explanatory variables. Moreover, in contrast to previous models and plot based approaches, they also allow for direct measurement of threshold effects.[2] Estimated dose-response curves in these new time series studies have for the most part indicated the presence of a linear relationship with no clear evidence of a threshold even at the lowest observed concentration levels.

[1] In one of the first of such analyses of threshold effects, Ostro (1984), used regression analysis to estimate linear spline exposure-response functions for London above the then regulatory limit of 50 $\mu g/m^3$. Examining data for London winters between 1958/9 and 1971/72, he found mortality effects at levels much lower than this level. Schwartz and Marcus (1990) used various autoregressive models to estimate mortality effects in the same data set and found a curvilinear relationship between mortality and particulate matter with steeper slopes occurring at lower particulate levels than higher ones.

[2] Single-city studies include: Fairly (1990), Dockery et al. (1992), Pope et al. (1992), Schwartz (1991, 1993, 1994), Schwartz & Dockery (1992a, b) and Xu et al. (2000). Multiple-city approaches include: Schwartz & Zanobetti (2000), Daniels et al. (2000) and Schwartz et al. (2001), Dominici et al. (2002).

In one of the most comprehensive of such studies, Dominici et al. (2002) estimate relative rates of daily mortality for PM and the shape of the dose-response curves for 88 of the largest U.S. cities. The authors estimate regional PM mortality dose-response curves using a hierarchical spline model that allows for the number and location of knots to vary from region to region while remaining the same across cities. The resulting curve was essentially linear with some regions indicating a slight departure from a linear model. Using data from 10 cities, Schwartz and Zanobetti (2000) use nonparametric smoothing functions to model the dose-response curve for daily mortality and PM and then combine results across all cities. Performing a number of simulations to confirm the ability of their approach to detect thresholds and other non-linear associations, the authors find no threshold levels. Schwartz et al. (2001) extend this nonparametric smoothing technique to allow for random variance components in order to estimate the shape of the dose-response curve for black smoke for 8 Spanish cities. The authors find that black smoke has a nearly linear relation with daily deaths but no threshold. This finding holds at even low pollution levels, and when other pollution and confounding factors are controlled for, and does not appear to be due to harvesting [Schwartz (2000a, 2000b, 2003)]. Schwartz et al. (2002) also fit log linear models using smoothing functions for $PM_{2.5}$. These nonparametric smoothing functions are estimated in a GAM that controls for temporal differences between cities. Log relative risks for exposure and the pointwise standard error for each city are then estimated for each exposure category. Estimates are combined to produce a national curve by regressing them against indicator variables at each exposure level in a model that also allows for city differences. The authors find no evidence of a threshold in either the single or combined dose-response relationship, with an essentially linear relationship holding between this pollutant and mortality.

The above studies are distinguished by the use of nonparametric smoothing approaches for estimation of the shape of the dose-response relationship. To illustrate the basic idea of previous approaches, the model of Dominici et al

(2002) can be written (ignoring the hierarchical aspect) as:

$$\log\left[E\left(y_t\right)\right] = f\left(PM_{t-1}\right) + X_t'\beta,$$

where $E\left(y_t\right)$ is expected mortality at time $t$ and $X_t$ are other explanatory variables. The effect of air pollution on mortality is measured through the unknown function $f\left(.\right)$ which may have nonlinearities such as threshold effects. There are many (related) ways of estimating $f\left(.\right)$ without assuming a particular parametric form for this function. Dominici et al (2002) use spline methods but other nonparametric smoothing methods are also used in this literature. The advantage of such an approach is that the researcher can let the data decide whether threshold effects (or any other nonlinearities) exist and at what level they occur. The disadvantage is that only nonlinearities in the $PM_{t-1}$-mortality relationship is allowed for. Put another way, this conventional approach is ideally designed to pick up threshold effects of the form: "$PM_{t-1}$ has no effect on mortality *when yesterday's PM levels were low*, but does have an effect if they were high". However, if the variable triggering the threshold effect is not $PM_{t-1}$, then the conventional approach will work poorly. For instance, the conventional approach will have trouble identifying threshold effects of the form: "$PM_{t-1}$ has no effect on mortality *when yesterday's ozone levels were low*, but does have an effect if they were high" or "$PM_{t-1}$ has no effect on mortality *when PM levels were steady over the past two days*, but does have an effect if they have been rising rapidly". Even those papers which give a nonparametric treatment to several explanatory variables (e.g. through a GAM), will not surmount this basic problem.

In contrast to this dominant approach, we introduce new parametric techniques to measure this relationship. We argue that our approach has benefits that make it more attractive than existing nonparametric approaches. On a general level, these advantages can be described as follows. Nonparametric methods are useful at tracing out the relationship between a dependent variable and an

explanatory variable when the exact form of the relationship is unknown. However, they must make assumptions about the smoothness of the relationship between air pollution and mortality. While they can accommodate any functional form (including thresholds), this extra flexibility may result in imprecise estimation. In contrast, parametric approaches have the advantage that they can focus flexibility in the area of interest (e.g. thresholds). This characteristic can potentially allow for more precise estimation. Second, nonparametric methods are poor when estimation involves many explanatory variables, all of which receive a nonparametric treatment (the so-called "curse of dimensionality"). High dimensionality is especially a problem in analyses of air pollution health effects, where many different pollutants, confounding factors and their interactions may play a significant role. To a certain extent, studies can avoid the problem by using GAMS, i.e. making the model additive. However, additivity does not come without a cost. The functional form must be assumed to be additive and hence interaction effects are not allowed for, an assumption that may not apply to air pollution mortality relationships.

Third, and most importantly in respect to air pollution health effects, as we have argued above, the question of which threshold triggers are involved in mortality effects is of great importance. For example, are adverse health effects triggered if the level of the pollutant is high on a single day? On two or more successive days? Does it or they have an effect only when pollution rises quickly? Is the average level of pollutant(s) the most important factor(s)? Perhaps the threshold effect is only evident in various interactions; for example, on hot days and when pollution rises rapidly? Ideally, a researcher will want to include as many potential threshold triggers as physical and health concerns may indicate, but most mortality-pollution studies look at only one such trigger: the level of pollution on a single day (usually yesterday). Using our approach, the researcher can incorporate a myriad of potential threshold triggers in order to find out which if any is important.

In this paper, we develop a model that allows for a myriad of different thresh-

old triggers. For a given threshold trigger, we have a regression model with a large number of potential explanatory variables (i.e. pollutants, meteorological variables and their lags as well as a spline to control for long term trends unrelated to air pollution). Different models are defined by different threshold definitions and the inclusion of different explanatory variables and, thus, our framework allows for a huge number of models. As argued below [see also Clyde (2000) and Koop and Tole (2004)], with so many models it is important to address the issue of model uncertainty through the use of Bayesian model averaging (BMA). In our empirical work, involving a data set from Toronto, Canada, we find that once we account for model uncertainty, the health effects of air pollution tend to be small and imprecisely estimated. Thus, the conclusion that air pollution has no effect on mortality (at least at the levels found in our data set) cannot be ruled out, even when threshold effects are allowed for. The only exception to this pattern is ozone, for which results are ambiguous. When $O_3$ is included with other pollutants as explanatory variables, we find no significant health effects or thresholds. However, when it is the only pollutant included as an explanatory variable, we find a significant mortality effect if ozone levels have been increasing above a threshold amount.

# 2 The Model

We have daily time series data for periods $t = 1, .., T$. Let $y_t$ denote the dependent variable (number of deaths) and we divide the explanatory variables into three categories: i) $x_{tj}^p$: the current day's level of pollutant $j$ for $j = 1, .., k_p$, ii) $x_{tj}^m$: the current day's value for $j = 1, .., k_m$ meteorological variables and iii) $x_t^o$: a row vector containing other explanatory variables (in our case these come from the spline used to control for long term trends and other systematic variations in mortality that may be unrelated to air pollution).

A standard model that does not allow for thresholds may be written as:

$$
\begin{aligned}
y_t &= \alpha + \sum_{i=0}^{p} x_{t-i,1}^{p}\beta_{i1} + + \sum_{i=0}^{p} x_{t-i,k_p}^{p}\beta_{ik_p} + \sum_{i=0}^{p} x_{t-i,1}^{m}\delta_{i1} + ... \qquad (1) \\
&\quad + \sum_{i=0}^{p} x_{t-i,k_m}^{m}\delta_{ik_m} + x_t^o\theta + \varepsilon_t,
\end{aligned}
$$

where $\varepsilon_t$ is i.i.d $N\left(0,\sigma^2\right)$. The health effects of the $j^{th}$ pollutant are measured by $\beta_{(j)} = \left(\beta_{0j},..,\beta_{pj}\right)'$. So, for instance, the measure we use, $\sum_{i=0}^{p}\beta_{ji}$, is the cumulative effect of the $j^{th}$ pollutant on mortality.[3]

If the explanatory variable triggering the threshold effects is known, then simple extensions of (1) can be developed to investigate threshold behavior. For instance, with a single pollutant, a (parametric) extension of (1) having threshold effects similar in spirit to the (semi-parametric) approach of Dominici, Daniels, Zeger and Samet (2002) will have:

$$
y_t = \begin{cases} \alpha^{(1)} + x_{t-1}^{p}\beta_1^{(1)} + \sum_{i=0}^{p} x_{t-i}^{m}\delta_{i1} + ... + \sum_{i=0}^{p} x_{t-i}^{m}\delta_{ik_m} \\ +x_t^o\lambda + \varepsilon_t \text{ if } x_{t-1}^{p} < r \\ \alpha^{(2)} + x_{t-1}^{p}\beta_1^{(2)} + \sum_{i=0}^{p} x_{t-i}^{m}\delta_{i1} + ... + \sum_{i=0}^{p} x_{t-i}^{m}\delta_{ik_m} \\ +x_t^o\lambda + \varepsilon_t \text{ if } x_{t-1}^{p} \geq r \end{cases} . \qquad (2)
$$

Thus, the effect of pollution on mortality is either$\beta_1^{(1)}$ or $\beta_1^{(2)}$ depending on whether yesterday's air pollution level, $x_{t-1}^{p}$, is below or above a threshold level, $r$. However, it is possible that yesterday's pollution level is not the threshold trigger. It may be that it is $x_{t-i}^{p}$ for $i \neq 1$. Or it may be that steep rises in air pollution may trigger health effects and thus $\Delta x_{t-i}^{p}$ is the threshold trigger. Likewise, cumulative build-up above some threshold may be key. In short, many threshold triggers may exist.

In light of these considerations, we extend (2) to:

[3]Note that we assume Normal errors throughout. Other work with daily mortality counts often uses Poisson regression methods, but their use would add greatly to the computational burden (unless approximations were used). Our total mortality data has mean 46.9, standard deviation 9.2, minimum 23, maximum 82 and a histogram that is roughly Normal. These considerations suggest that the costs associated with working with Normal errors are small (i.e. our dependent variable takes on many values and its discrete distribution can be very well approximated by a continuous Normal distribution). These costs are likely to be much greater than those associated with the use of approximate Poisson methods.

$$y_t = \begin{cases} \alpha^{(1)} + \sum_{i=0}^{p} x_{t-i,1}^{p} \beta_{i,1}^{(1)} + \sum_{i=0}^{p} x_{t-i,k_p}^{p} \beta_{i,k_p}^{(1)} + \sum_{i=0}^{p} x_{t-i}^{m} \delta_{i1}^{(1)} + ... \\ \qquad + \sum_{i=0}^{p} x_{t-i}^{m} \delta_{ik_m}^{(2)} + x_t^o \lambda + \varepsilon_t \text{ if } z_{t-d} < r \\ \alpha^{(2)} + \sum_{i=0}^{p} x_{t-i,1}^{p} \beta_{i,1}^{(2)} + \sum_{i=0}^{p} x_{t-i,k_p}^{p} \beta_{i,k_p}^{(2)} + \sum_{i=0}^{p} x_{t-i}^{m} \delta_{i1}^{(2)} + ... \\ \qquad + \sum_{i=0}^{p} x_{t-i}^{m} \delta_{ik_m}^{(2)} + x_t^o \lambda + \varepsilon_t \text{ if } z_{t-d} \geq r \end{cases} . \quad (3)$$

Thus, the threshold effect is triggered by an explanatory variable, $z_t$, with a delay of $d$ days.[4]

In the existing literature, semiparametric or spline-based approaches to smoothing are common . These have the advantage that they can allow for a wide range of nonlinearities in functional form. But they also have the disadvantage that they focus on the nonlinear effects of a single explanatory variable. As discussed previously, a spline approach to estimating the unknown function $f\left(x_{t-1}^{p}\right)$ will work very well threshold effects exist and $x_{t-1}^{p}$ is the threshold trigger. However, it will work poorly if an explanatory variable other than $x_{t-1}^{p}$ is the threshold trigger. Our parametric approach, which can accommodate many potential threshold triggers, will work well in both cases. Of course, it may not work well in the presence of other types of nonlinearities. But given the importance of threshold effects to policymakers, a parametric model that focuses on threshold issues in a very general and flexible way has advantages over alternative approaches (which are possibly flexible in many directions of little interest, but may not be flexible enough in the threshold-effect direction).

In this paper, we allow for $z_t$ to depend on various possible transformations of the pollutant. This yields the following threshold triggers:

1. $z_{t-d} = x_{t-d}^{p}$ for $d = 0, .., q$.
2. $z_{t-d} = \frac{\sum_{j=0}^{d} x^{p}{}_{t-d}}{d+1}$ for $d = 1, .., q$.
3. $z_{t-d} = \Delta x_{t-d}^{p}$ for $d = 0, .., q-1$.
4. $z_{t-d} = \frac{x^{p}{}_{t} - x^{p}{}_{t-d}}{d}$ for $d = 2, .., q$.

[4] In the previous equations, we have assumed the same number of lags for every explanatory variable. This assumption can easily be relaxed. However, as we shall see in our discussion of Bayesian model averaging, $p$ is a maximum lag length and our estimation procedure considers models with shorter lag lengths that may differ across explanatory variables.

The first of these threshold triggers allows levels of variables up to $q$ days ago to be the threshold trigger, while the second allows for average levels over $d$ days. The third allows for changes in variables to trigger threshold effects, while the fourth depends on average changes over the last $d$ days. All of these threshold triggers are plausible in the context of an air pollution-mortality study. The first is the standard definition but the second is also plausible if only consistently high pollution levels over several days trigger health effects. The third and fourth definitions relate to growth rates in air pollution and, if rapid changes in air pollution levels trigger mortality effects, these may be the reasonable definitions. Note that the fourth definition is good at capturing a cumulative buildup of air pollutants as a trigger in threshold effects.[5]

To avoid confusion, we remind the reader of our terminology. We refer to $z_t$ as the "threshold trigger" that depends on the "delay" ($d$) and the "threshold" ($r$). Note, too, that the model given by (3) can be extended in many directions. It is trivial to extend our framework such that $z_t$ depends on any explanatory variable for which the researcher has data, $w_t$. In this paper, we set $w_t = x_t^p$, thus allowing for threshold effects triggered by the pollutant. However, $w_t = x_{tj}^m$ or $w_t = x_{tj}^m x_t^p$ for $j = 1, .., k_w$ can easily be accommodated. Furthermore, $z_{t-d}$ may be a vector of several possible triggers. Likewise, $z_{t-d}$ can depend on unknown parameters (e.g. an unknown weighted average of pollution levels over several days). Or we could allow for several thresholds (e.g. equation 3 could be augmented with a third equation and we could allow for air pollution to have different mortality effects depending on whether $z_{t-d}$ was high, medium or low). In this paper, we do not consider such extensions as our model is already very flexible and risks over-parameterization. Furthermore, our computer programs, although simple to write, take a long time to run. Dealing with such extensions risks turning a computationally demanding project into a computationally infeasible one.

[5] To avoid reundant or degenerate definitions we allow the delay parameter, $d$, to vary across definitions.

# 3 Overview of Econometric Methods

In previous work, we argued for the use of Bayesian model averaging in the context of air pollution studies [Koop and Tole (2004)]. In that paper, we remained within the linear framework of (1). We argued that the need to include many pollutants and meteorological variables (and their lags) as well as a spline to control for trends in mortality unrelated to air pollution means that such studies will naturally include a large number of potential explanatory variables. With large numbers of potential explanatory variables, serious problems associated with over-parameterization may occur [see the discussion and references in Koop and Tole (2004)]. For our purposes, it is sufficient to note that it is not sensible to include all potential explanatory variables (many of which may end up having no explanatory power) in one large regression. Including irrelevant variables reduces the accuracy of any estimation procedure. This problem has motivated the use of sequential hypothesis testing as a way of eliminating apparently irrelevant explanatory variables. The problems associated with the presentation of results from a single model selected on the basis of a sequence of hypothesis tests have long been recognized in the statistical literature. Intuitively, each time a hypothesis test is carried out, a possibility exists that a mistake has been made (i.e. the researcher will reject the better model for a not so good one). This possibility multiplies sequentially with each hypothesis test. Second, even if a sequential hypothesis testing procedure does lead to the selection of the best model, decision theoretical arguments imply that it is rarely desirable to present results for this model while ignoring all evidence from the not quite so good model(s).

Bayesian model averaging is a common way of surmounting problems associated with selection of a single model. Koop and Tole (2004) and others [e.g. Clyde (2000)] discuss the benefits of Bayesian model averaging (both theoretically and for empirical practice) in air pollution-mortality studies. Results are based on a weighted average of results from many models. The weights in

the average are based on the probability that each model is the correct one. The basic idea behind Bayesian model averaging can be summarized as follows: Suppose a researcher is entertaining $R$ possible models, denoted by $M_1, ..., M_R$, in order to learn about a parameter of interest, $\theta$ (e.g. the effect of a pollutant on health). Since it is not known which model is correct, models are treated as random variables. The posterior model probability, $p(M_r|Data)$ summarizes what we know about the models after seeing the data. The rules of conditional expectation imply that:

$$E(\theta|Data) = \sum_{r=1}^{R} p(M_r|Data) E(\theta|Data, M_r).$$

In words, the overall point estimate of $\theta$ is the weighted average of the point estimates in every model. The weights in the weighted average are the posterior model probabilities,[6] $p(M_r|Data)$ for $r = 1, .., R$.[7] The reader interested in more justification for and discussion of Bayesian model averaging is referred to: Draper (1995), Fernandez, Ley, and Steel (2001a, b), Hodges (1987), Hoeting, Madigan, Raftery and Volinsky (1999) and Raftery, Madigan, and Hoeting (1997).

In general, if we have $K$ potential explanatory variables and wish to consider all models defined by inclusion/exclusion of every potential explanatory variable then we must deal with $2^K$ models. Even in the standard Normal regression model (without thresholds), if $K$ is at all large (e.g. $K > 20$) it is computationally difficult and even impossible to obtain results for every possible model. This motivates our use of a computationally more efficient algorithm referred to as Markov Chain Monte Carlo Model Composition (or MC$^3$) originally developed by Madigan and York (1995). More detail is provided in the Technical

[6] For the non-Bayesian, the posterior model probability is proportional to the marginal likelihood. The log of the marginal likelihood is, asymptotically, closely related to common information criteria (e.g. the Schwarz or Hannan-Quinn criteria). Hence crude intuition implies that Bayesian model selection is comparable to model choice using an information criterion and Bayesian model averaging is comparable to averaging across models using weights proportional to (the exponential of) the information criterion.

[7] This equation can be extended to any function of the parameters, $f(\theta)$. For instance, setting $f(\theta) = \theta^2$ will allow for the calculation of $E(\theta^2|Data)$, which is needed to calculate the posterior standard deviation.

Appendix. However, we note that even with this algorithm, the computational demands of our approach are very high and limit the number of possible threshold definitions we can consider.

Problems arising from a large number of potential explanatory variables are exacerbated by the need to control for long term trends and other systematic variations in mortality that may be unrelated to air pollution. Splines are typically used to correct for such effects. Recent statistical work on the use of splines in air pollution-mortality studies [e.g. Clyde (2000)] makes three main conclusions: i) including a spline is potentially important; ii) the precise choice of class of spline (e.g. cubic, thin plate, etc.) is relatively unimportant; iii) the precise choice of time scale (i.e. the number of knots) is potentially very important. In respect to the latter, if we include too few knots, we do not fully correct for the unknown trend terms (e.g. the increase in mortality caused by flu epidemics could be attributed to air pollution). However, if we include too many knots, then important health effects may be removed (i.e. the spline will be so flexible as to explain all the variation in mortality, leaving nothing left for air pollution to explain). Thus, we follow a commonly-used strategy [see, e.g., Clyde (2000) or Smith and Kohn (1996)]. We choose a particular class of spline, put in numerous knots, and then use Bayesian model averaging to deal with the excessive number of explanatory variables.

In this paper, we use a thin plate spline with a knot placed every 60 days. If we let $n_j$ denote the knot at time $j$ and $N$ the number of knots, then the unknown trend is given by:

$$f(t) = \alpha_0 + \sum_{j=1}^{N} \alpha_j b_j(t),$$

where

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

From a statistical point of view, the key point to note is that $b_j(t)$ can be

interpreted as an explanatory variable. In terms of (3), the explanatory variables implied by the spline are included in $x_t^o$.

For a given threshold variable, (3) contains $2 \times (k_w + k_p)(p+1) + N + 1$ potential explanatory variables.[8] In our empirical work, we have $k_w = 4$, $p = 3$, $N = 36$ and $k_p = 1$ or 2. Thus, we have either 77 or 85 potential explanatory variables. However, we also have a large number of possible threshold triggers. We choose a maximum delay parameter of $q = p$ and, thus, have eight possible choices for $z_{t-d}$. We consider three possible thresholds corresponding to unusually low, average and unusually high levels of the threshold trigger (i.e. we set $r = -1, 0, 1$ standard deviations of $z_{t-d}$). Thus, we use 36 (for the single pollutant) or 72 (for the case with two pollutants) models of the form of (3). Our empirical work involves a huge BMA exercise averaging over all potential explanatory variables for all threshold trigger/threshold choices.

The statistical methods used to carry out Bayesian inference (including Bayesian model averaging) are described in the Technical Appendix. Here we include only a brief summary. Let $\phi$ be a 4-vector to denote which of the four types of threshold definition is used such that $\phi_j = 1$ if the $j^{th}$ definition is used ($\phi_j = 0$ otherwise) and let $\omega = \left(\phi', r, d\right)'$. Conditional upon a choice for $\omega$, (3) can be written as one big Normal linear regression model (i.e. all explanatory variables, including the thresholds, are known and the two equations are written as one using appropriately defined dummy variables). Let $K$ denote the number of potential explanatory variables in this big model and $\gamma = (\gamma_1, .., \gamma_K)'$ be a vector of 0-1 random variables denoting which of the potential explanatory variables is contained in a particular model. So, for instance, $\gamma = (1, .., 1)'$ is the big model containing all the potential explanatory variables, $\gamma = (1, .., 1, 0)'$ contains all explanatory variables with the exception of the last one, etc.. As described above, there are $2^K$ possible configurations of $\gamma$. Finally, let $\psi = \left(\alpha', \beta', \delta', \lambda', \sigma^2\right)'$ denote all the other parameters of the model.

We treat $\omega, \gamma$ and $\psi$ as unknown parameters and develop a posterior sim-

[8] An intercept is included in all models and thus is not included in the calculation of this total.

ulation algorithm that allows for Bayesian inference. This algorithm draws from $p(\gamma|Data, \omega)$ for a particular configuration of $\omega$. Standard methods for Bayesian model averaging [Koop and Tole (2004)] can be adapted to draw from $p(\gamma|Data, \omega)$. At each draw, appropriate functions of $p(\psi|Data, \omega, \gamma)$ are evaluated (e.g. $E(\psi|Data, \omega, \gamma)$). Assuming a natural conjugate prior, $p(\psi|Data, \omega, \gamma)$ has the Normal-Gamma form. These draws are then averaged to provide $E(\psi|Data, \omega)$ and other relevant functions of $\psi$. Standard textbook results (based on the marginal likelihood for the Normal linear regression model with natural conjugate prior) provide a formulae for $p(\omega|Data, \gamma)$ [e.g. Koop (2003)]. Thus, $p(\omega|Data, \gamma)$ can be evaluated at every draw for $\gamma$. Averaging over all draws provides an estimate that is proportional to $p(\omega|Data)$.

In this way, BMA results can be obtained for a given choice of $\omega$ and $p(\omega|Data)$ can be evaluated for that choice. By repeating the steps outlined in the previous paragraph for every possible configuration of $\omega$ we obtain BMA results that average over all possible explanatory variable choices for all threshold definitions.

Although we advocate the use of Bayesian model averaging, it is worthwhile noting that the posterior simulator can also be used in Bayesian model selection. This can be done by finding the $\gamma, \omega$ combination or the configuration of $\omega$ with the highest probability.

# 4 Empirical Results

This section presents empirical results from daily time series data for Metropolitan Toronto for the years 1992-1997. The complete data set is described in the Data Appendix. Our explanatory variables include those from the spline described in the previous section, four meteorological variables (i.e. $x_t^m$ in equation 3 contains measures of barometric pressure, temperature, relative humidity and cloud) and pollutants. Given the serious computational demands of Bayesian model averaging in such a framework, we limit the number of pollutants to one or two. When working with two pollutants, we use $O_3$ and PM as having the

greatest potential importance based on recent studies and policy concern. When working with one pollutant, one approach we adopt is to calculate a Pollution Index. To do this, we take our data on $SO_2$, NO, $NO_2$, CO, $O_3$ and PM and use principal component methods to extract the first factor.[9] We also present results using $O_3$ and PM individually.

The class of models is defined by (3) with definitions of the variables that might potentially trigger threshold effects given in Section 2. We choose a maximum lag length of three days (i.e. $p = 3$), which is enough to capture any short term effects (and is at least as large as used in other studies in the literature). For this reason, we set the maximum delay in the threshold effect to be three days as well (i.e. $q = 3$). For the thresholds themselves, we use three possible values corresponding to threshold changes occurring at unusually low, average, and unusually high levels of the threshold trigger. In terms of (3), this means that we normalize all threshold triggers (i.e. $z_t$) by subtracting off their means and dividing by standard deviations and then consider $r = -1, 0, 1$. Working with two pollutants, implies 72 different potential threshold triggers whereas one pollutant implies 36.

The decision about which empirical results to present also depends on an interpretational issue. Recall that the key features of interest in this study are measures of the effect of pollution on mortality *above and below a threshold level for a certain variable we have called the threshold trigger*. In our study, both the threshold level and the threshold trigger vary across models and, thus, the concepts implied by the phrase "the effects of the pollutant above and below the threshold" differ across models. Put another way, one model will produce "the effect of $O_3$ on mortality, if $O_3$ levels yesterday were above average" while another model will produce "the effect of $O_3$ on mortality, if the average levels of $O_3$ over the past three days were unusually high". These are conceptually quite different and it does not make sense to simply average over both of them when presenting evidence relating to the effect of pollution in a high threshold regime.

[9] The first factor accounts for 85.3% of the total variability in the pollutants.

Hence it also does not make sense to carry out Bayesian model averaging over the thresholds.

In light of these issues (and to present as wide a view of the data as possible), we present two sorts of empirical results. First, we present traditional Bayesian model selection results based on the one single model with maximum posterior model probability (i.e. one single threshold definition and one single choice of explanatory variables). In terms of the notation of Section 3, we find the values for $\gamma$ and $\omega$ that yield the highest posterior model probability. Secondly, we present results that combine aspects of Bayesian model averaging and Bayesian model selection. In particular, we choose the single threshold definition that has the highest posterior probability and carry out BMA using this definition (and all potential explanatory variables). Thus, we find the value for $\omega$ that yields the highest posterior model probability, but carry out BMA over $\gamma$. We also present the probability that the particular threshold chosen is the correct one (so that the reader may, if desired, weight our results by this probability in an informal Bayesian model averaging exercise). For comparative purposes, we also present results for models which do not include any thresholds at all.

In our empirical results, we refer to the effect of a pollutant on mortality. We remind the reader that this effect is the cumulative effect (i.e. in terms of equation 3, $\sum_{i=0}^{p} \beta_{ji}^{(1)}$ and $\sum_{i=0}^{p} \beta_{ji}^{(2)}$ are the effects below and above the threshold). Since our explanatory variables are all normalized, a point estimate of, say, 0.5 can be interpreted as saying: "the cumulative effect of a one standard deviation increase in the pollutant, maintained for $p$ days, is an extra 0.5 of a death".[10]

[10] Many researchers present mortality effects of air pollution in terms of a 10 unit change (e.g. 10 $\mu$g m$^{-3}$ for the case of particulate matter). We prefer to express our results as effects of one standard deviation changes since standard deviations have the same interpretation for all pollutants. For readers interested in translating our results: One standard deviation implies 9.15 ppb for $O_3$ and 12.12 $\mu g$ per m$^3$ for PM.

| Table 1: Estimate of Effect of Pollutant on Mortality with Two Pollutants: $O_3$ and PM<br>Estimate is Posterior Mean (Posterior standard deviation in parentheses) | | | | | |
|---|---|---|---|---|---|
| Pollutant | Effect if $z_{t-d} < r$ | Effect if $z_{t-d} \geq r$ | Choice for $z_{t-d}$ | Choice for $r$ | Prob. Selected Model is Correct |
| Bayesian Model Selection for No-threshold case, choosing $\gamma$ (No BMA) | | | | | |
| $O_3$ | 0.516<br>(0.175) | 0.516<br>(0.175) | n.a. | n.a. | 0.002 |
| PM | 0.000<br>(0.000) | 0.000<br>(0.000) | n.a. | n.a. | 0.002 |
| BMA for No-threshold case | | | | | |
| $O_3$ | 0.276<br>(0.300) | 0.276<br>(0.300) | n.a | n.a | n.a. |
| PM | 0.203<br>(0.260) | 0.203<br>(0.260) | n.a. | n.a. | n.a. |
| Bayesian Model selection, choosing $\gamma$ and $\omega$ (No BMA) | | | | | |
| $O_3$ | 0.000<br>(0.000) | 0.000<br>(0.000) | PM<br>$d = 2$ | $r = -1$ | 0.001 |
| PM | 0.000<br>(0.000) | 0.000<br>(0.000) | PM<br>$d = 2$ | $r = -1$ | 0.001 |
| Mixed Bayesian Model Selection (choosing $\omega$) and BMA (for $\gamma$) | | | | | |
| $O_3$ | 0.241<br>(0.309) | 0.291<br>(0.305) | PM<br>$d = 2$ | $r = -1$ | 0.326 |
| PM | 0.051<br>(0.157) | 0.100<br>(0.235) | PM<br>$d = 2$ | $r = -1$ | 0.326 |

| Table 2: Estimate of Effect of Pollutant on Mortality with Single Pollution Index<br>Estimate is Posterior Mean (Posterior standard deviation in parentheses) | | | | | |
|---|---|---|---|---|---|
| Pollutant | Effect if $z_{t-d} < r$ | Effect if $z_{t-d} \geq r$ | Choice for $z_{t-d}$ | Choice for $r$ | Prob. Selected Model is Correct |
| Bayesian Model Selection for No-threshold case, choosing $\gamma$ (No BMA) | | | | | |
| Index | 0.000<br>(0.000) | 0.000<br>(0.000) | n.a. | n.a. | 0.003 |
| BMA for No-threshold case | | | | | |
| Index | 0.277<br>(0.285) | 0.277<br>(0.285) | n.a | n.a | n.a. |
| Bayesian Model selection, choosing $\gamma$ and $\omega$ (No BMA) | | | | | |
| Index | 0.000<br>(0.000) | 0.000<br>(0.000) | Index<br>$d = 0$ | $r = -1$ | $9.9\times10^{-5}$ |
| Mixed Bayesian Model Selection (choosing $\omega$) and BMA (for $\gamma$) | | | | | |
| Index | 0.204<br>(0.263) | 0.228<br>(0.308) | $\Delta$Index<br>$d = 0$ | $r = 1$ | 0.094 |

| Table 3: Estimate of Effect of Particulate Matter on Mortality<br>Estimate is Posterior Mean (Posterior standard deviation in parentheses) | | | | | |
|---|---|---|---|---|---|
| Pollutant | Effect if $z_{t-d} < r$ | Effect if $z_{t-d} \geq r$ | Choice for $z_{t-d}$ | Choice for $r$ | Prob. Selected Model is Correct |
| Bayesian Model Selection for No-threshold case, choosing $\gamma$ (No BMA) | | | | | |
| PM | 0.000<br>(0.000) | 0.000<br>(0.000) | n.a. | n.a. | 0.003 |
| BMA for No-threshold case | | | | | |
| PM | 0.238<br>(0.267) | 0.238<br>(0.267) | n.a | n.a | n.a. |
| Bayesian Model selection, choosing $\gamma$ and $\omega$ (No BMA) | | | | | |
| PM | 0.000<br>(0.000) | 0.000<br>(0.000) | PM<br>$d = 2$ | $r = -1$ | 0.002 |
| Mixed Bayesian Model Selection (choosing $\omega$) and BMA (for $\gamma$) | | | | | |
| PM | 0.067<br>(0.191) | 0.274<br>(0.270) | PM<br>$d = 2$ | $r = -1$ | 0.573 |

| Table 4: Estimate of Effect of Ozone on Mortality<br>Estimate is Posterior Mean (Posterior standard deviation in parentheses) | | | | | |
|---|---|---|---|---|---|
| Pollutant | Effect if $z_{t-d} < r$ | Effect if $z_{t-d} \geq r$ | Choice for $z_{t-d}$ | Choice for $r$ | Prob. Selected Model is Correct |
| Bayesian Model Selection for No-threshold case, choosing $\gamma$ (No BMA) | | | | | |
| $O_3$ | 0.516<br>(0.175) | 0.516<br>(0.175) | n.a. | n.a. | 0.004 |
| BMA for No-threshold case | | | | | |
| $O_3$ | 0.317<br>(0.308) | 0.317<br>(0.308) | n.a | n.a. | n.a. |
| Bayesian Model selection, choosing $\gamma$ and $\omega$ (No BMA) | | | | | |
| $O_3$ | 0.000<br>(0.000) | 0.900<br>(0.242) | $\Delta O_3$<br>$d = 2$ | $r = 0$ | 0.007 |
| Mixed Bayesian Model Selection (choosing $\omega$) and BMA (for $\gamma$) | | | | | |
| $O_3$ | 0.162<br>(0.283) | 0.780<br>(0.362) | $\Delta O_3$<br>$d = 2$ | $r = 0$ | 0.406 |

In our previous work with linear models, Koop and Tole (2004), we found that point estimates (i.e. posterior means) of the mortality effects of air pollution were consistently positive (indicating pollution was bad for health), but that they were all small relative to their posterior standard deviations (i.e. the Bayesian analysis to the standard error).[11] In fact, we found that a mortality effect of air pollution of zero was always plausible. The overall message of Tables 1 through 4 is similar (with one important exception). In addition, there

[11] For the non-Bayesian reader, note that the posterior standard deviation is analogous to the standard error. We adopt the commonly used rule-of-thumb that a posterior mean that is two standard deviations from zero is "significant".

is little evidence of threshold effects in the air pollution-mortality relationship (with one important exception). We expand on these points in the remainder of this section.

Table 1 contains results using two pollutants, $O_3$ and PM. It is very interesting to note, if we do not allow for thresholds and do not carry out BMA, that $O_3$ appears to have a substantive (and significant) effect on mortality (i.e. the posterior mean of the effect is 0.516, which is almost three times as large as its standard deviation). However, when we carry out BMA and, thus, allow for a proper treatment of model uncertainty, the point effect becomes smaller and is not significant. That is, the point estimate is 0.276, which is less than one posterior standard deviation from zero. This strengthens the case, emphasized in our previous work, that model selection procedures that ignore model uncertainty can lead to seriously misleading results. Furthermore, even when we carry out Bayesian model selection, allowing for threshold effects removes the apparent importance of $O_3$ for mortality. In fact, the point estimate of 0.000 implies that the single most probable model does not even include $O_3$ as an explanatory variable.

In Table 1, PM has no appreciable effect on mortality irrespective of whether Bayesian model averaging is performed and whether thresholds are included.

Tables 2 and 3 include results from single pollutant cases, using our Pollution Index (constructed using six individual pollutants) and PM, respectively. They reveal a familiar pattern: point estimates of the health effects of air pollution are positive or zero, but are all small relative to posterior standard deviations that take into account model uncertainty. This statement holds irrespective of whether thresholds are included.

Tables 2 and 3 thus indicate that no strong threshold effects of any sort exist in the pollution-mortality relationships we examine. Nevertheless, it is interesting to see which threshold definitions are favored by the data and, in particular, how the preferred choices vary both across pollutants and statistical methodology. For instance, in Table 1, when we select a single threshold

and then carry out BMA over all potential explanatory variables, our preferred threshold trigger is the *level* of PM two days ago and we obtain $r = -1$ (i.e. it is at unusually *low* levels of PM that the threshold occurs). The comparable BMA results in Table 2 imply that the preferred threshold trigger is the *growth rate* of pollution over the current day and $r = 1$ (i.e. it is at unusually *high* rates of pollution growth that the threshold occurs). However, if we carry out Bayesian model selection, Table 2 implies it is unusually low *levels* of pollution today that trigger the threshold. In short, there is an enormous degree of uncertainty over which is the appropriate threshold trigger. This is reflected in the the last columns of the tables. For instance, Tables 1 and 2 indicate that there is only a 38.3% and 9.4% chance, respectively, that the chosen threshold is the correct one. Thus, by choosing a single threshold choice, we ignore other choices with a 62.7% and 90.6% probability of being correct.

Table 4 contains the one exception to these findings: ozone. The model for this pollutant without thresholds reveals the same patterns as in Table 1. That is, if we simply select the most probable model, we find $O_3$ to have a positive and significant (i.e. more than two posterior standard deviations from zero) effect on mortality. Yet, when we perform BMA in this no-threshold case, the effect vanishes. However, when we allow for thresholds the effect of $O_3$ on mortality is positive and significant. This table implies a clear pattern: Below a threshold, $O_3$ has no effect on mortality, but above the threshold $O_3$ has a positive effect. The positive effect kicks in if above average changes in ozone two days ago (a different threshold trigger than in other studies) occur. However, it is worth stressing that this significant effect occurs with a threshold trigger which is only 40.6% likely to be the appropriate definition.

The results in Table 1 and Table 4 are apparently in contradiction with one another. Table 1 indicates that there are no significant mortality effects of ozone, whereas Table 4 indicates a positive and significant effect above a threshold. The reason for this apparent contradiction is that, in our mixed Bayesian model selection and BMA approach we first choose a threshold definition and then

present BMA results for that definition. This definition differs between Tables 1 and 4. In Table 1, the level of PM two days ago is the threshold trigger; in Table 4 it is the growth rate of ozone two days ago. It is interesting to note that, in Table 1, the most probable threshold definition receives 32.6% probability. The second most probable threshold definition (with 23.5% of the probability) is the same as that in Table 4. If we perform BMA using both pollutants and this second most probable threshold definition, the point estimate of the effect of $O_3$ on health in the above threshold regime is positive and almost significant (i.e. posterior mean 0.711, standard deviation 0.389). Hence, if we had gone with this second most probable threshold choice, results in Table 1 and Table 4 would have been quite similar.[12]

These results may indicate some more complicated role for PM and ozone, perhaps due to complex interactions between ozone and PM arising from gas to particle conversion processes [Meng et al. (1997)]. Alternatively, it may simply be due to the model containing only ozone suffering from the omitted variables bias. Another possibility may be multicollinearity; that is, a strong correlation between PM and ozone that makes it impossible to disentangle the individual effects of each. However, the correlation between the two is only .21, which indicates multicollinearity is not a problem.

What overall conclusion should the reader take from these empirical results? First, with the exception of ozone, for no pollutant do we find a threshold effect. We have argued that model uncertainty is an important issue in any air pollution modelling strategy. Hence our use of a Bayesian model averaging approach that presents measures of uncertainty (e.g. posterior standard deviations) reflecting our ignorance about which model is the true one. When we perform BMA we are unable to find point estimates that are "significant" in a conventional sense. The second most important message concerns the pollutant ozone, for which results are ambiguous. Ozone appears to be quite sensitive to the type of pollutants

[12] Note that our modeling strategy also allows for threshold effects in the impacts of meteorological variables. For the sake of brevity, we do not focus on these impacts here. However, it is worthwhile noting that they do explain why some threshold definitions are preferred by the data even if they imply no threshold effects in the impacts of the pollutants on mortality.

included in the analysis. If we include both ozone and particulate matter as explanatory variables (and, thus, measure the effect of ozone *after controlling for* PM) we find no significant threshold effects (unless we work with threshold definitions that are not preferred by the data). If we include only ozone as an explanatory variable (and, thus, measure the effect of ozone *without controlling for* PM) we do find significant threshold effects. Which result is more compelling will partly depend on which set of controls the reader is interested in. A third message concerning ozone is that threshold effects, when found to exist, seem to be triggered by unconventional variables. It is common for researchers [e.g. Dominici, Daniels, Zeger and Samet (2002)] to investigate the use of yesterday's pollution level as a threshold trigger. In no case have we found this to be the most plausible threshold trigger. This suggests that our modelling strategy, which allows for a myriad of possible threshold triggers, is a useful alternative to nonparametric approaches that depend on a single threshold trigger.

Finally, as with any air pollution-mortality study using daily time series data, it is important to stress that our findings do not necessarily imply that air pollution does not have adverse health effects. Rather, our results indicate that there is no reliable statistical evidence for a link between air pollution and mortality in the particular daily time series data set that we consider. It is possible that health effects short of mortality (e.g. asthma attacks) are caused by air pollution. Furthermore, studies involving daily time series data will be unable to pick up any long term health effects of air pollution. Nevertheless, given that studies with daily time series data are widely used to set air pollution standards in the U.S. and elsewhere in the world, we argue that it is important that researchers use appropriate statistical methods to estimate air pollution impacts. We hope that our modeling strategy is a step in this direction.

# 5 References

Clyde, M. (2000). Model uncertainty and health effect studies for particulate matter. *Environmetrics,* 11, 745-763.

Daniels, M.J., Dominici, F., Samet, J.M., Zeger, S.L. (2000). Estimating particulate matter mortality dose-response curves and threshold levels: an analysis of daily time series for the 20 largest U.S. cities. *American Journal of Epidemiology*, 152, 397-406.

Delfino, R., Becklake, M., Hanley, J. and Singh, B. (1994). Estimation of unmeasured particulate air pollution data for an epidemiological study of daily respiratory morbidity. *Environmental Research,* 67, 20-38.

Dockery, D.W., Schwartz, J. and Spengler, J.D. (1992). Air pollution and daily mortality: associations with particulates and acid aerosols. *Environmental Research.* 59, 362-373.

Dominici, F., Daniels, M., Zeger, S.L. and Samet, J.M. (2002). Air pollution and mortality: estimating regional and national dose-response relationships. *Journal of the American Statistical Association*, 97, 100-111.

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

Fairly, D. (1990). The relationship of daily mortality to suspended particulates in Santa Clara County, 1980-1986. *Environmental Health Perspectives,* 89, 159-168.

Fernandez, C., Ley, E. and Steel, M. (2001a). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16, 563-576.

Fernandez, C., Ley, E. and Steel, M. (2001b). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100, 381-427.

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

Hoeting, J., Madigan, D., Raftery, A. and Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14, 382-417.

Koop, G. (2003). *Bayesian Econometrics.* Chichester: John Wiley and Sons.

Koop, G. and Tole, L. (2004). Measuring the health effects of air pollution: To what extent can we really say that people are dying from bad air? *Journal of Environmental Economics and Management*, 47, 30-54.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review,* 63, 215-232.

Meng, Z., Dabdub, D. and Seinfeld, J.H. (1997). Chemical coupling between atmospheric ozone and particulate matter. *Science*, 277, 116-119.

Ostro, B. (1984). A search for a threshold in the relationship of air pollution to mortality: A reanalysis of data on London winter. *Environmental Health Perspectives*, 58, 397-399

Pope, C.A. 3rd., Schwartz, J. and Ransom, R.R. (1992). Daily mortality and PM pollution in Utah Valley. *Archives of Environmental Health,* 47, 211-217.

Raftery, A., Madigan, D. and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92, 179-191.

Schwartz. J., and Dockery, D.W. (1992a). Particulate air pollution and daily mortality in Steubenville, Ohio. *American Journal of Epidemiology*, 135, 12-19.

Schwartz, J., and Dockery, D.W. (1992b). Increased mortality in Philadelphia associated with daily air pollution concentrations. *American Review of Respiratory Disorders*, 145, 600-604.

Schwartz, J. and Marcus, A. (1990). Mortality and air pollution in London: a time series analysis. *American Journal of Epidemiology.* 131, 185-194.

Schwartz, J. (1991). Particulate air pollution and daily mortality in Detroit. *Environmental Research*, 56, 204-213.

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

Schwartz, J. (1994). Total suspended particulate matter and daily mortality in Cincinnati, Ohio. *Environmental Health Perspectives,* 102, 186-189.

Schwartz, J. (2000a). Assessing confounding, effect modification, and thresholds in the association between ambient particles and daily deaths. *Environmental Health Perspectives,* 108, 563-568.

Schwartz, J. (2000b). Is there harvesting in the association of airborne

particles with daily deaths and hospital admissions? *Epidemiology*, 12, 55-61.

Schwartz, J. (2003). Air pollution and daily mortality in a city with low levels of pollution. *Environmental Health Perspectives*, 111, 45-51.

Schwartz, J., Ballester, F., Saez, M., Perz-Hoyos, S., Bellido, J., Cambra, K., Arribas, F., Canada, A., Perez-Boillos, M.J. and Sunyer, J. (2001). The concentration-response relation between air pollution and daily deaths. *Environmental Health Perspectives,* 109, 1001-1006.

Schwartz, J., Laden, F., and Zanobetti, A. (2002). The concentration-response relation between $PM_{2.5}$ and daily deaths. *Environmental Health Perspectives*, 110, 1025-1029.

Schwartz, J. and Zanobetti, A. (2000). Using meta-smoothing to estimate dose-response trends across multiple studies, with application to air pollution and daily death. *Epidemiology,* 11, 666-672.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75, 317-343.

Xu, Z., Yu, D. and Jing., L. (2000). Air pollution and daily mortality in Shenyang, China. *Archives of Environmental Health*, 47, 115-20.

# 6 Data Appendix

In this study, we use daily data on mortality, pollutants and meteorological variables from 1992-1997. Each variable is discussed below. The chosen time span was dictated by the fact that mortality data was only available through 1997 and regular collection of data on some of the key pollutants only began in 1992.

**Mortality Data**

The mortality data was provided by the Toronto Department of Public Health and covers all deaths in the Metro Toronto area (i.e. the municipalities of Toronto, Etobicoke, York, North York, East York and Scarborough). The data contain total daily deaths and deaths by various disease categories. Of these, we use the variables: total deaths, deaths due to diseases of the circu-

latory system and deaths due to diseases of the respiratory system. For reasons of confidentiality, if the number of deaths in any disease category is between 1 and 4 the precise value is not reported. In our data, this suppression of information only occurs with deaths due to diseases of the respiratory system and helps motivate our focus on total mortality. When we ran our programs using respiratory deaths, we coded all suppressed values as the average of 1 and 4 (i.e. 2.5). We do not report these results here since they were similar to those found using total mortality.

**Weather Data**

Hourly data on the following climate variables was provided by Ontario Climate Centre at Environment Canada from their Pearson International Airport monitoring station:

- Pressure (0.01 kilopascals).
- Temperature (0.1 degrees C).
- Relative humidity (%).
- Total cloud amount (tenths).
- Precipitation (0.1 mm). Note: this variable is a daily total.
- Visibility (0.1 km).
- Wind direction (10s of degrees). Note: this variable is transformed into a 1-8 scale in the standard way [see Delfino et al. (1995), page 22).
- Wind speed (km per hour).

The first four of these are used directly in our regressions, the last four are used only in the regression to fill in missing values for particulate matter (see below). There are very few missing values in the hourly data. Absent values are replaced by an average of values for the hours before and after the missing value. To create daily data from this hourly data, we take a daily mean. Empirical results using daily maxima are very similar.

**Pollution Data**

Hourly data on the following air pollution variables was provided by the Air Monitoring Section of the Ontario Ministry of Environment:

- $SO_2$ (ppb).
- NO (ppb).
- $NO_2$ (ppb).
- NOX (ppb).
- COH = coefficient of haze (0.1 COH/1,000 ft.)
- CO (ppm).
- $O_3$ (ppb).

We average data from the six monitors that have nearly complete data for 1992-1997. These monitors are widely dispersed across various street locations in Metro Toronto: In downtown Toronto (Bay/Grosvenor), Scarborough (Lawrence/Kennedy), North York (Yonge/Finch), Etobicoke (Elmcrest), Etobicoke (Evans/Arnold) and York (Clearview/Keele). Missing values are handled in the same manner as the weather variables. There are relatively few missing values. The worst monitor had 2% of its hourly observations missing. To create daily data from this hourly data, we simply take a daily mean. Empirical results using daily maxima are very similar.

Daily averages for airborne particulate matter were provided by the Analysis and Air Quality Division of Environment Canada. Fine particulate matter ($PM_{2.5}$) is defined as having a size less than 2.5 micrograms. Coarse particulate matter ($PM_{10-2.5}$) is defined as being between 2.5 and 10 micrograms in diameter. Total particulate matter, PM, is the sum of fine and coarse particulate matter. For the years 1992-1994 the only available monitor was at Bay/Wellesley streets. For 1996-1997 the only available monitor was at Evans/Arnold streets.

For 1995, data from both monitors were available. This overlap year was used to correct for the small difference in means between the two monitors.

Missing values are a serious problem in most studies that include particulate matter. For example, the standard approach in the U.S. is to sample every sixth day. Our data set is of better quality, providing roughly one observation every three days. Nevertheless, 66.29% of our raw daily observations are missing. In order to provide estimates of the missing observations, we follow a procedure similar to that described in Delfino et al. (1994). In particular, we use the non-missing particulate matter values to run a regression using relevant explanatory variables. We then use the values of the explanatory variables on the missing days and the estimated regression coefficients to predict particulate matter values for days for which data are missing. Following Delfino et al. (1994), we use daily means and maximums of all the pollution and weather variables listed above as explanatory variables. Delfino et al. (1994) suggests a particular nonlinear transformation of some of the key variables, but we find that simply adding squares of all explanatory variables provides a better fit. For $PM_{2.5}$ the resulting regression has an $R^2$ of 0.72 while for $PM_{10-2.5}$ the $R^2$ is 0.50. Note that the resulting fitted values for the particulate matter data contain information from other pollutants and weather variables. However, most of the explanatory power comes from variables that are not included in the mortality regressions. In particular, visibility, wind direction and the coefficient of haze provide most of the explanatory power in the regressions in which particulate matter variables are the dependent variables.

# 7 Technical Appendix

Our posterior simulator involves sequentially drawing from $p(\gamma|Data, \omega)$, $p(\omega|Data, \gamma)$ and $p(\psi|Data, \omega, \gamma)$ where these parameter vectors are described in the body of the paper.

We have data for $t = 1, .., T$ days[13] and denote data on the dependent

[13] When $p$ lags are included in the model, we proceed conditionally upon $p$ initial observations

variable (mortality) by $y = (y_1, .., y_T)'$. All the potential explanatory variables including lags and including the threshold variables (see equation 3) are stacked in a $T \times K$ matrix $X(\omega)$. We stress that the notation $X(\omega)$ implies that the explanatory variables depend on the threshold triggers. Models are defined, using $\gamma$, as containing different subsets of all the explanatory variables in $X(\omega)$. For a given choice of $\omega$ we have $r = 1, .., R$ models, denoted by $M_r$ or, equivalently, by $\gamma^{(r)}$. The explanatory variables in $M_r$ are denoted by $X(\omega)_r$. Thus we have

$$y = \alpha\iota_T + X_r(\omega)\beta_r + \varepsilon \tag{A.1}$$

where $\iota_T$ is a $T \times 1$ vector of ones , $X_r(\omega)$ is a $T \times k_r$ matrix containing some (or all) columns of $X$. The $T-$vector of errors, $\varepsilon$, is assumed to be $N\left(0_T, \sigma^2 I_T\right)$ where $0_T$ is a $T-$vector of zeros and $I_T$ is the $T \times T$ identity matrix. Note that we are assuming all models contain an intercept.

We use a Normal-Gamma natural conjugate prior with hyperparameters chosen in the objective fashion described in Fernandez, Ley and Steel (2001b). To be precise, for the error variance we use the standard noninformative prior:

$$p(\sigma) \propto \frac{1}{\sigma}. \tag{A.2}$$

We standardize all the explanatory variables by subtracting off their means and dividing by their standard deviations. Once this is done, it makes sense to use a flat prior for the intercept:

$$p(\alpha) \propto 1. \tag{A.3}$$

For the slope coefficients we assume a g-prior of the form:

$$\beta_r \sim N\left(0_{k_r}, \sigma^2\left[g_r X_r(\omega)' X_r(\omega)\right]^{-1}\right). \tag{A.4}$$

and hence $y_1$ will actually be the $p^{th}$ day of January, 1992.

It remains only to specify $g_r$. Fernandez, Ley and Steel (2001b) investigate the properties for many possible choices for $g_r$, including values that lead to posterior model probabilities with properties similar to commonly-used information criteria (e.g. the Schwarz or Hannan-Quinn criteria). Their recommendation is to choose:

$$g_r = \begin{cases} \frac{1}{K^2} \text{ if } T \leq K^2 \\ \frac{1}{T} \text{ if } T > K^2 \end{cases} .$$

Our empirical application uses this choice for $g_r$, although other considered choices lead to qualitatively similar results.

The resulting posterior for $\psi$ (i.e. the regression coefficients and error precision) conditional upon $\omega$ follows a multivariate t-distribution with mean, covariance and degrees of freedom parameters being of standard textbook form [see, e.g., Poirier (1995) or Koop and Tole (2004)].

The posterior model probability for model $r$ in the Bayesian model averaging is:

$$\begin{aligned} p\left(M_r|Data,\omega\right) &\equiv p\left(\gamma^{(r)}|Data,\omega\right) \\ &= c\left(\frac{g_r}{g_r+1}\right)^{\frac{k_r}{2}} \left[ \begin{array}{c} \frac{1}{g_r+1} y' P_{X_r} y + \\ \frac{g_r}{g_r+1}\left(y-\overline{y}\iota_T\right)'\left(y-\overline{y}\iota_T\right) \end{array} \right]^{-\frac{T-1}{2}}, \end{aligned} \tag{A.5}$$

where

$$P_{X_r} = I_T - X_r\left(\omega\right)\left[X_r\left(\omega\right)' X_r\left(\omega\right)\right]^{-1} X_r\left(\omega\right)' .$$

and $c$ is a constant that is the same for all models. The fact that $\sum_{r=1}^{R} p\left(M_r|Data,\omega\right) = 1$ can be used to evaluate $c$ (if required).

To draw $\gamma^{(r)}$ from (A.5), we adopt the MC$^3$ described in Madigan and York (1995). This Metropolis algorithm is very simple to implement. In particular, if the current model in the chain is $\gamma^{(s)}$ then a candidate model, $\gamma^{(j)}$, that is randomly (with equal probability) selected from the set of models including

$\gamma^{(s)}$ and which change $\gamma^{(s)}$ by either switching a 1 to a 0 or a 0 to a 1 (i.e. the algorithm either randomly adds or subtracts one column from $X_s(\omega)$), is drawn. $\gamma^{(j)}$ is accepted with probability:

$$\min\left\{1, \frac{p\left(\gamma^{(j)}|Data\right)}{p\left(\gamma^{(s)}|Data\right)}\right\}. \tag{A.6}$$

If $\gamma^{(j)}$ is not accepted then the chain stays with $\gamma^{(s)}$. It can be shown that the relative frequency that each model is drawn will converge to its posterior model probability.

Posterior results based on the sequence of models generated from the MC$^3$ algorithm can be calculated by averaging over draws of $\gamma$. That is, if $S$ draws of $\gamma$ are taken (after discarding initial burn-in draws), then for any function $g(.)$, $E[g(\psi)|Data, \omega]$ can be approximated by $\widehat{g}_S$ where

$$\widehat{g}_S = \frac{1}{S}\sum_{s=S_0+1}^{S} E\left[g(\psi)|Data, \gamma^{(s)}, \omega\right]. \tag{A.7}$$

BMA results (e.g. posterior moments of the cumulative effect of the pollutant) can be obtained by running this MC$^3$ algorithm for all possible configuration of $\omega$ and then taking a weighted average with weights given by $p(\omega|Data)$. To evaluate $p(\omega|Data)$ we note that all elements of $\omega$ are discrete random variables which take on a finite number of values. We allow for $r$ to take on three possible values: $-1, 0, 1$. Since $z_{t-d}$ is standardized by subtracting off its mean and dividing by its standard deviation, these values for $r$ have a simple interpretation common to all threshold triggers. Assuming a noninformative Uniform prior for each element of $\omega$, $p(\omega|Data, \gamma)$ is proportional to the marginal likelihood for the Normal linear regression model defined by $\omega$ and $\gamma$. Hence, by evaluating the marginal likelihood for every configuration of $\omega$ (at each value for $\gamma$) we can average (in a manner analogous to A.7) to estimate $p(\omega|Data)$.

To monitor convergence of the chain we calculate the probability of the ten most probable models drawn in two different ways. First, we calculate them

analytically using (A.5). Then we approximate this probability using output from the $MC^3$ algorithm. When these probabilities to three decimal places are the same, we deem convergence to have taken place. The number of draws required for the various models considered were typically around 1,000,000 per threshold definition.