Measuring the Health Costs of Air Pollution: To What Extent Can We Really Say that People are Dying from Bad Air?

Gary Koop
Department of Economics
University of Glasgow
G.Koop@socsci.gla.ac.uk

and

Lise Tole
Centre for Development Studies
University of Glasgow
ltole@socsci.gla.ac.uk
and
The World Bank, Washington, D.C.
ltole@worldbank.org

July 2002
Revised: November 2002

ABSTRACT: The estimation of the costs of environmental impacts is a major focus of current theoretical and policy research in environmental economics. Such estimates are commonly used, for example, to set regulatory standards for pollution exposure, to design appropriate environmental protection and damage mitigation strategies, to guide the assessment of environmental impacts, and to measure public willingness to pay for environmental amenities. It is a truism to say that the effectiveness of such strategies depends crucially on the quality of the estimates used to inform them. However, this paper argues that in respect to at least one area of the empirical literature – the estimation of the health impacts of air pollution – most existing cost estimates may be questionable and thus have limited relevance for environmental decision-making. By neglecting the issue of model uncertainty – that is, which models, among the myriad of possible models researchers should choose from to estimate costs – most studies overstate confidence in their chosen model and underestimate the evidence from rejected models, thereby greatly enhancing the risk of obtaining uncertain and inaccurate results. This paper discusses the importance of model uncertainty for accurate estimation of the health costs of air pollution. Its importance is demonstrated in an exercise that models pollution-mortality impacts using a new and comprehensive data set for Toronto, Canada. The main empirical finding of the paper is that standard deviations based on commonly used point estimates for the measurement of air pollution-mortality impacts become very large when model uncertainty is incorporated into the analysis. Indeed, they become so large as to question the plausibility of previously measured links between air pollution and mortality. Although applied to the estimation of air pollution costs, the general message of this paper – that proper treatment of model uncertainty critically determines the accuracy of the resulting estimates – applies to many studies that seeks to estimate environmental costs.

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 express our sincere thanks to Janet Phillips, Toronto Department of Public Health, for her expert assistance with the mortality data and for her valuable responses to our many requests for information. We are also grateful to the editor of this journal, two anonymous referees and Mingliang Li for many useful comments.

1 Introduction

The estimation of both the indirect and direct costs of environmental degradation motivates a great deal of current theoretical and policy research in environmental economics. Indeed, the fact that pollution is perceived to have large impacts that can be quantified in economic and human terms informs much of the theoretical and policy literature on pollution regulation, abatement, and assessment of associated environmental and health impacts.[1] The applicability of these estimates to the formulation of effective environmental policies hinges crucially on the magnitude of the measured costs of the observed impact. A wide variety of empirical techniques have been used to estimate the costs of pollution, including time series and cohort studies of health effects and contingent valuation studies of the willingness to pay for pollution. At the real-world level of policy-making, studies that measure the health impacts of air pollution currently inform air pollution standards worldwide. For example, reacting to a recent U.S. Supreme Court ruling (February 27, 2001) that upheld the responsibility of the EPA to "set air quality standards at the level that is 'requisite' – that is, not lower or higher than is necessary – to protect the public health with an adequate margin of safety" – the Executive Director of the Clean Air Trust summarized the link between science and policy thus: "For more than 30 years, the bedrock principle of the Clean Air Act is that national clean air standards should be based on science and the impact on public health. This bedrock principle has driven three decades of clean air progress" (ENS 2002).

However, if empirical estimates of costs are fragile or uncertain, then the basic assumptions underlying large parts of the environmental economics literature are called into question. This paper will argue that – at least in respect to one area of this literature – the estimation of human health effects of air pollution – it is quite difficult to estimate these costs precisely once it is acknowledged that there are many plausible models that could be used to estimate them. Our results apply to a particular data set and a particular environmental cost (i.e. the health effects of air pollution). However, our general message – that proper treatment of model uncertainty is crucial – holds for virtually any study that seeks to estimate environmental costs.

This paper attempts to estimate the health effects of air pollution in a large metropolitan city using time series data. The time series literature on air pollution-mortality effects, which uses various measures of mortality, pollutants and meteorological variables, has tended to find that air quality does have an effect on mortality (see, among many others, Burnett et al. 1997, Dominici et al. 2000, Ross et al. 1996, Ostro,

[1] This literature is too voluminous to cite here. A few recent contributions include Alberini et al. (1997), Delucchi, Murphy and McCubbin (2002), Hahn (2000), Khanna and Damon (1999), Krupnick et al. (1999, 2000), Navrud (2001) and Schennach (2000).

Hurley and Lipsett, 1999, Schwarz, 1993). However, concerns have been raised in the statistical literature about whether these findings are an artifact of data mining (i.e. of presenting results from a single model based on sequential testing procedures). A 1998 report by the U.S. National Research Council also made the investigation of this issue a research priority (see National Research Council, 1998). On a related issue, there is disagreement over exactly which confounding variables to include in the analysis of pollution health impacts (e.g. meteorological variables, time effects that capture long term trends, other temporal factors due to flu epidemics, and so on). The multiplicity of potential models raises questions about the appropriate model with which to estimate these impacts.

An increasingly popular way of surmounting the problems associated with sequential hypothesis testing procedures is to use Bayesian model averaging techniques. A series of papers, Clyde (2000), Clyde and DeSimone-Sasinowska (1997) and Clyde, Guttorp and Sullivan (2000), use Bayesian model averaging procedures to investigate the effect of particulate matter on mortality. These papers find that model uncertainty is a very important determinant of results, at least for data sets for Phoenix, Arizona and Birmingham, Alabama. Of particular importance are the findings that posterior distributions for relative risks are fairly dispersed and allocate appreciable probability in regions near one. In other words, these studies find that the hypothesis that particulate matter has no effect on health is not so unlikely. Proper treatment of model uncertainty is the fundamental message of this literature; indeed, we argue that the credibility of the finding that pollution may result in death may hinge on this issue.

The use of Bayesian model averaging in the measurement of the health effects of air pollution distinguishes our work from virtually all of the related literature. The main exception is that by Clyde (2000), Clyde and DeSimone-Sasinowska (1997) and Clyde, Guttorp and Sullivan (2000). The present study differs from these papers in several ways. Chiefly, we use a more extensive data set on a hithero largely unanalyzed city: Toronto, Canada. We also use a wide range of pollutants rather than just one. Furthermore, the work of Clyde et al. does not include potentially important interactions between pollutants or between pollutants and meteorological variables. In contrast, the comprehensive nature of our data set and the inclusion of interaction terms are an advantage of the present study. These advantages, however, cause computational problems in that the huge number of explanatory variables mean that a direct implementation of traditional Bayesian model averaging algorithms is impossible. Consequently, we suggest various ways of addressing this problem.

We find that our empirical results are robust to choice of particular algorithm. That is, we find that point estimates of health effects of pollutants are positive. Yet, when we allow for model uncertainty, we find that standard deviations become so large that the hypothesis of no effect is always plausible. We caution that this finding does not necessarily mean that air pollution has no effect on human health. Rather, as we argue, a thorough econometric analysis, which treats the problems raised by model uncertainty, indicates that there is not enough information in the time series data to estimate its effect on mortality precisely, at least for our chosen case study of Toronto.

2 Bayesian Model Averaging

In theory, estimates of the health effects of air pollution can be found by running a regression of a health variable (e.g. mortality) on relevant pollutants and other explanatory variables (e.g. meteorological variables). This is what most studies do. However, in practice, there is uncertainty over which pollutants and which meteorological variables are relevant for the study. Hence the researcher may wish to investigate many possible explanatory variables. Furthermore, the precise timing of health effects is unclear, and for this reason, many lags of the explanatory variables should be included. Another important issue (which has not been extensively explored in the literature) is whether important interactions exist between explanatory variables (e.g. whether health effects worsen if both ozone and particulate matter are high on a given day). In short, an enormous number of potential explanatory variables exist, and, likewise, an enormous number of potential models. If there are K potential explanatory variables and each model is defined by the inclusion or omission of an explanatory variable, then there are 2^K possible models. In the present application, K could easily be 100 or more, implying billions of possible models.[2] The usual practice is to use hypothesis testing procedures to select a single model from among the billions of potential models, and to present results from this model as being representative of the "true" model.

The problems associated with the presentation of results from a single model selected on the basis of a sequence of hypothesis tests has long been recognized in the statistical literature. Increasingly, these problems are being acknowledged in applied economics (e.g. Fernandez, Ley and Steel, 2001a and Sala-i-Martin, 1997). Statistical discussions of these problems are provided in many places. For instance, Poirier (1995), pp. 519-523 provides a theoretical discussion of the problems with pre-test estimators. Draper (1995) and Hodges (1987) are also important references in this field.

For the purposes of this paper, we need only provide a brief, intuitive, description of key issues which these papers address. First, these papers draw attention to the fact that each time a hypothesis test is carried out, the possibility always exists that a mistake will be made (i.e. the researcher will reject the "better" model

[2] Since 2^{100} is more than 10^{30} even referring to "billions of models" is possibly a misleading underestimate.

for a “not so good” one). The possibility of making a mistake quickly multiplies as sequences of hypothesis tests are carried out. So, for instance, a claim that a regression t-statistic of 2.0 means that a hypothesis is rejected at the 5% level of significance is spurious and, potentially vastly misleading, if the regression is selected on the basis of previous hypothesis tests. Second, even if a sequential hypothesis testing procedure does lead to the selection of the “best” model, standard decision theory implies that it is rarely desirable to simply present results for this model and ignore all evidence from the “not quite so good” model(s). Generally speaking, this is reflected in the common empirical wisdom that, if you mine the data long enough you are bound to find something – but you should not put too much trust in your finding.

Given problems caused by sequential hypothesis testing procedures, the researcher may be tempted simply to include all potential variables in a regression. However, this approach is also unsatisfactory since including irrelevant variables tends to decrease the accuracy of the estimation, making it difficult to uncover effects that may really exist. In classical statistical procedures, including irrelevant explanatory variables will increase standard errors, making it difficult to find significant effects. These pitfalls motivate an increasing interest among researchers in model averaging: a method in which empirical 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. However, formally classical econometric methods do not allow for the calculation of the “probability that a model is the correct one”. For this reason, many have turned to Bayesian methods. The literature on Bayesian model averaging has burgeoned in recent years (see Hoeting et al, 1999, for a recent survey and practical guide). The basic idea behind Bayesian model averaging can be explained quite simply: Suppose the researcher is entertaining R possible models, denoted by $M_1, ..., M_R$, to learn about a parameter of interest, θ (e.g. the effect of a pollutant on health). For the models considered in this paper, it is straightforward to use the data to calculate the probability that a model is a correct one. That is, $p\left(M_r|Data\right)$ can be calculated for $r = 1, .., R$. It is also straightforward to calculate a point estimate of θ in every model. We take the posterior mean, $E\left(\theta|Data, M_r\right)$, as this point estimate. According to the rules of probability, it follows that:

$$E\left(\theta|Data\right) = \sum_{r=1}^{R} p\left(M_r|Data\right) E\left(\theta|Data, M_r\right). \tag{2.1}$$

In words, the overall point estimate of θ is the weighted average of the point estimates in every model. The weights in the weighted average are the posterior model probabilities, $p\left(M_r|Data\right)$ for $r = 1, .., R$. This same logic applies to functions of θ so, for instance, we can use:

$$E\left(\theta^2|Data\right) = \sum_{r=1}^{R} p\left(M_r|Data\right) E\left(\theta^2|Data, M_r\right) \tag{2.2}$$

to help us calculate the posterior variance of θ, which can then be used to quantify uncertainty about θ.[3] The precise formulae for $p\left(M_r|Data\right)$ and $E\left(\theta|Data, M_r\right)$ are provided in the Technical Appendix to this paper. To provide some intuition we note that $E\left(\theta|Data, M_r\right)$ is similar to an OLS estimate and $p\left(M_r|Data\right)$ shares some similarities with information criteria such as the Schwarz criteria or Akaike information criteria.

Four additional points should be stressed at this stage. First, many researchers feel that the real world is very complicated and that all models under consideration are likely to be approximations of reality and, thus, wrong. If all models under consideration are wrong, then model averaging can be interpreted as a way of adding robustness to protect against misleading inferences. Second, Bayesian methods allow for the incorporation of prior information about the parameters of the model. However, in this paper we do not elicit such a prior but rather use the objective or benchmark prior recommended in Fernandez, Ley and Steel (2001b). Third, with the enormous number of models under consideration, it is not possible to evaluate $p\left(M_r|Data\right)$ and $E\left(\theta|Data, M_r\right)$ for every model. This is a common occurrence in empirical work involving Bayesian model averaging. A literature has developed that devises various ways of overcoming the problem. In this paper we use an algorithm described in Madigan and York (1995) referred to as Markov chain Monte Carlo model composition (MC^3). Intuitively, this involves randomly drawing models in such a way that a given model is drawn with frequency proportional to $p\left(M_r|Data\right)$. In this way, the algorithm focuses on the models with high probability (which thus receive high weight in the model averaging procedure), avoiding the models with low probability. Further details are given in the Technical Appendix. Fourth, for the non-Bayesian reader, it should be stressed that the different between our results and those obtained using traditional methods are due to the treatment of model uncertainty and not due to any other aspects of the Bayesian methodology. For instance, in this paper, we use priors which are fairly noninformative relative to the data. Numerous empirical and theoretical papers have shown that, in the context of a single model and without strong prior information, Bayesian and non-Bayesian studies will yield yield similar point estimates (i.e. posterior means and OLS estimates will be similar) and measures of parameter uncertainty (i.e. posterior standard deviations and standard errors will be similar). It is likely that a classical statistical methodology which used model averaging would lead to similar results as those presented here.[4]

[3] To be precise, the posterior variance of θ is given by $var\left(\theta|Data\right) = E\left(\theta^2|Data\right) - \left[E\left(\theta|Data\right)\right]^2$.

[4] Formally, model averaging cannot be done in a classical framework since models are not random variables.Hence, for the classical econometrician statements such as "the probability that a model is true" has no well-defined meaning. Model averaging can, however, be carried out classically in an *ad hoc* fashion using, e.g., penalized likelihoods or information criteria to weight different models (see Sala-i-Martin, 1997).

3 Empirical Results

3.1 Overview

This section contains empirical results using daily time series data from Metropolitan Toronto for the years 1992-1997. The complete data set is described in the Data Appendix. Given our message – that empirical results should reflect model uncertainty – we do not simply present our final results. Instead, we offer some discussion of and motivation for the route that leads us to our final specification.

In all cases, our dependent variable is a measure of mortality. For reasons discussed in the Data Appendix, we focus on total mortality although we note that findings for deaths due to diseases of the circulatory and respiratory systems are very similar.[5] A myriad of potential explanatory variables could be included here. Typically, previous researchers have chosen a small subset of relevant variables, focusing on a single pollutant and/or using hypothesis testing procedures to discard many potential variables from the analysis. In the previous section, we argued that such a procedure could lead researchers to make seriously misleading inferences about the health effects of air pollution. Consequently, we use a much bigger set of explanatory variables involving seven pollutants and five weather variables. The seven pollutants are denoted by SO_2, CO, NO, NO_2, O_3, $PM_{2.5}$ and $PM_{2.5-10}$.[6] All of these have been used (usually one at a time) in previous studies, although recently interest has focused on fine and coarse particulate matter ($PM_{2.5}$ and $PM_{2.5-10}$). The study's five weather variables are barometric pressure (PRESSURE), temperature (TEMP), humidity (HUMIDITY), total amount of cloud (CLOUD) and wind speed (WIND).[7] All explanatory variables are standardized by subtracting their mean and dividing it by their standard deviation.[8] The Data Appendix provides further details about the data. The weather and pollution variables provide us with twelve explanatory variables. However, we recognize the possibility that there may be important interactions between these variables (e.g. the effects of pollutants may worsen on hot days or the effect of a certain pollutant may be greater when combined with high values of another pollutant). Thus, it is important to allow for interactions between all the weather and pollution variables. However, there are 66 possible interactions. Including each of the original variables along with every possible interaction between them gives us 78 explanatory variables.

[5] The units of our dependent variables are deaths per day. We note that results using logged daily deaths do not alter our basic empirical findings.

[6] We have data on NO_x, but this variable is nearly perfectly correlated with NO and, for this reason, it is excluded from the analysis.

[7] Schwarz (1993) does not include wind speed since it is unlikely that this variable directly influences mortality. However, low wind speed tends to be associated with high pollution events and, thus, there is a risk that the effects of pollution on mortality will be counfounded incorrectly with wind speed. In an early version of this paper, we eliminated wind speed as well. Results were virtually the same as those presented here.

[8] All explanatory variables are daily averages. Results using other transformations of the original hourly data (e.g. daily maxima) are essentially the same as those presented here.

In addition, it is important to allow for time lags of every variable. In this paper we include the current value of every explanatory variable along with up to three lags. This results in 312 potential explanatory variables.

However, there are still more potential explanatory variables that must be included in the analysis. In time series analyses of pollution-health effects, it it is obviously important to control for long term trends and other systematic variations in mortality that are unrelated to air pollution. Splines are typically used to correct for such effects.[9] In particular, spline methods are often used to approximate relationships of unknown form and, thus, are closely related to nonparametric methods. Simply put, a spline is a flexible way of approximating the relationship between two variables when the exact functional form is unknown. In contrast, a linear regression assumes a straight line relationship between two variables – y and x – such that the functional form is linear. However, it is highly unlikely that the relationship between mortality and outbreaks of the flu, for example, will conform to a straight line; experience tells us that it will more likely fluctuate, rising for a few weeks and then falling as infection rates decline. Hence the need for a method that does not make assumptions about functional form.

Recent statistical work on the use of splines in air pollution-mortality studies (see, 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). In light of these considerations, the recommended strategy, which we follow here, is to 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

[9] Spline methods are often used to approximate relationships of unknown form and, thus, are closely related to nonparametric methods. Our spline is a function of time. To fit a spline, the researcher chooses several points in time (referred to as knots) and fits the unknown trend by connecting the dots. Various curves can be used to connect the dots and the decision on what type of curve to fit (e.g. cubic, etc.) determines the class of splines.

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

From a statistical point of view, the key point to note is that $b_j(t)$ can be interpreted as an explanatory variable and α_j as a regression coefficient. Thus adding a spline is akin to adding explanatory variables to a regression. Thus, our Bayesian model averaging approach is not complicated by adding a spline. Nevertheless, the number of additional explanatory variables can be quite large. In our case, inclusion of a spline adds 36 explanatory variables.

Thus, a very general model would include 348 potential explanatory variables. That is, we have twelve air pollution and meteorological variables and 66 interactions between these variables. If we include current values and up to three lags of all variables the number rises to 312 potential explanatory variables. Adding in the 36 explanatory variables implied by the spline leads to 348 in total. However, directly implementing Bayesian model averaging for more than approximately 50 potential explanatory variables is impossible given current computational limitations. Consequently, we cannot directly use the algorithm outlined in the Technical Appendix on the full model. In the remainder of this section we describe various special cases or algorithm modifications that allow for the implementation of Bayesian model averaging. Our strategy is to investigate the air pollution-mortality relationship using a variety of different approaches on the grounds that empirical findings that are robust across various approaches are more reliable than results using one approach.

3.2 Case 1: No Interaction Terms and No Spline

We begin by presenting results for what might be considered a conventional case. Very few empirical studies have included interaction terms and many of the less statistically-sophisticated studies do not include splines (or other terms that control for trends in mortality). Thus a model that only includes our seven pollutants and five meteorological variables (and three lags of each) is a good starting point for our investigation of the mortality effects of air pollution. This type of reasoning implies a regression with 48 potential explanatory variables. Thus, as described in the Technical Appendix, conventional Bayesian model averaging, which directly uses the MC^3 algorithm, is computationally demanding but feasible

Table 1 presents the proportion of models visited by the MC^3 algorithm which contained each explanatory variable. Intuitively, these numbers can be interpreted as the probability that each explanatory variable has a substantive effect and should be included in the model. It can be seen that some of the meteorological variables undoubtedly have an effect on mortality (e.g. the current day's barometric pressure should be included with 98.9% probability and the temperature three days ago enters 82.9% of the models). However, none of the pollutants at any of the lags considered enters with any appreciable probability. Yesterday's level of carbon monoxide enters with 34.9% probability and the current day's level of ozone is included 14.2% of the time, but these probabilities are all quite low. Thus, it seems unclear whether any of the pollutants has an appreciable effect on mortality.

Table 1: Probability of Including Each Explanatory Variable

Explanatory Variable	Lag			
	0	1	2	3
Pollutants				
SO_2	0.032	0.027	0.031	0.067
CO	0.047	0.349	0.058	0.035
NO	0.023	0.042	0.046	0.044
NO_2	0.026	0.067	0.074	0.054
O_3	0.142	0.026	0.026	0.029
$PM_{2.5-10}$	0.021	0.024	0.023	0.020
$PM_{2.5}$	0.040	0.070	0.021	0.021
Meteorological Variables				
PRESSURE	0.989	0.497	0.415	0.141
TEMP	0.089	0.347	0.214	0.829
HUMIDITY	0.025	0.050	0.033	0.023
CLOUD	0.023	0.045	0.027	0.135
WIND	0.095	0.067	0.096	0.020

This finding is bolstered if we calculate the cumulative effect of each pollutant on health using Bayesian model averaging. This cumulative affect is the standard multiplier (i.e. for any pollutant, we sum the coefficients of the current value and all lags). Table 2 presents the posterior mean (i.e. a point estimate) and posterior standard deviation (i.e. a measure of uncertainty in the point estimate akin to a standard error) of the cumulative effect of each pollutant. To aid in interpretation, remember that the explanatory variables have been standardized so that a cumulative effect of, say, 0.5, for a particular pollutant, means that a rise in that pollutant of one standard deviation (maintained over four days) is associated with an additional 0.5 deaths.

Table 2: Cumulative Effect of Each Pollutant on Mortality

	Posterior Mean	Posterior Standard Deviation
SO_2	0.029	0.106
CO	0.200	0.260
NO	0.025	0.109
NO_2	0.056	0.156
O_3	0.054	0.159
$PM_{2.5-10}$	0.004	0.055
$PM_{2.5}$	0.029	0.111

As expected, the point estimates in Table 2 are all positive (indicating that air pollution is harmful to health). However, the magnitude of all of these effects is quite small and the posterior standard deviations are very large. Thus, there exists enormous uncertainty over the magnitude of the health effect of these pollutants. Put another way, faced with this statistical evidence, we expect no researcher would feel confident offering policy advice of the form: "The cumulative effect of fine particulate matter on mortality is 0.029". The uncertainty associated with this point estimate is much too large for such advice to be taken seriously.

The reason for why the posterior standard deviations are so large is that model uncertainty is huge. The ten most probable models in total only account for 12.2% percent of the total probability. And there is only a 3.9% chance that the best model is a correct one. Thus, any analysis which selects only the single best model will be basing inference on a model which is 96.1% sure to be incorrect! To illustrate the effect of model uncertainty, Figure 1 plots the posterior of the cumulative effect of ozone on mortality. The spike in the posterior at zero means that most of the probability is associated with models where ozone (and its lags) do not enter the model. That is, models which elicit statements of the form: "ozone has no effect on mortality" receive most support from the data. Another noteworthy point is that the posterior allocates some probability to negative values for the effect of ozone on mortality. In other words, some models exist that actually imply that ozone should be beneficial to health! Given our models are defined according to the inclusion or exclusion of explanatory variables, we can say that some (relatively implausible) combinations of explanatory variables imply that the effect of ozone on mortality is negative.

The most probable model includes only weather variables. However, the third most probable model includes CO lagged one period as an explanatory variable (as well as some weather variables). The Bayesian who estimated this single model using a noninformative prior would obtain a posterior mean of 0.405 and posterior standard deviation of 0.163 for the coefficient on CO lagged one period. The non-Bayesian using this single model would have obtained an OLS estimate of this coefficient of 0.405 and a t-statistic of 2.48. Thus, either of these econometricians who ran only this single regression would conclude that CO has a large, statistically significant effect on mortality. Given the imperfections associated with model selection techniques like stepwise regression, it is distinctly possible that a researcher could end up selecting this third most probable model and reporting strong health effects for CO. However, from our Bayesian model averaging perspective we can calculate that there is only a 1.8% chance that this model is the correct one. This illustrates how a method that presents results from a single regression has the potential to lead researchers to make misleading inferences about pollution-mortality effects, seriously underestimating the true uncertainty about the statistical evidence.

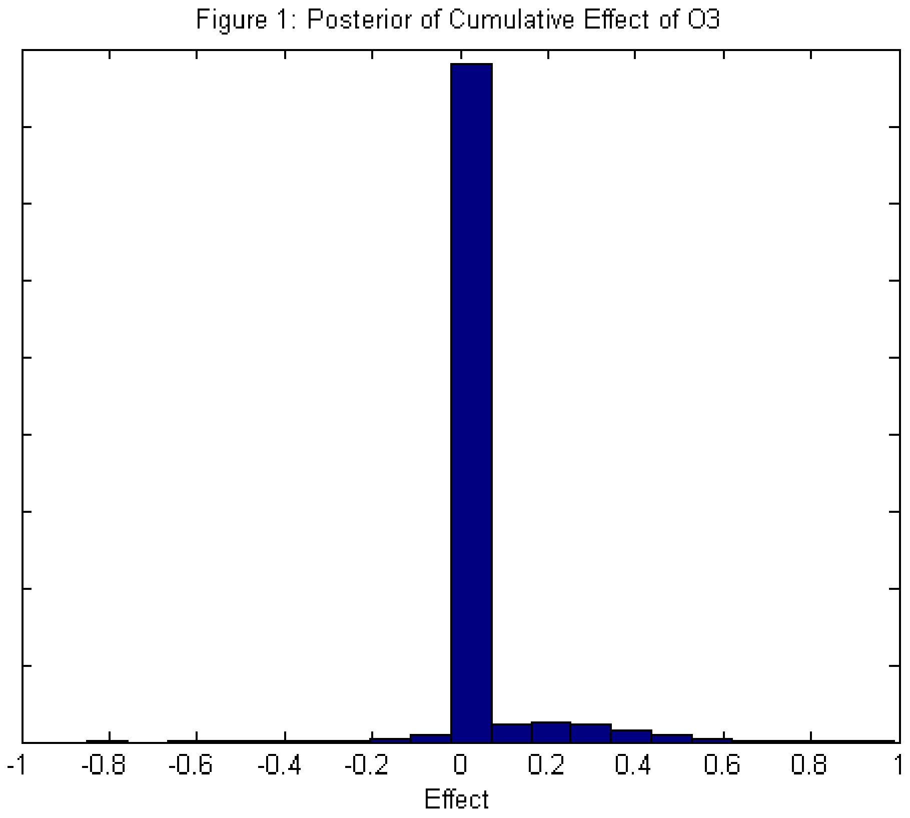
Figure 1: Posterior of Cumulative Effect of O3
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
Effect

Figure 1:

3.3 Case 2: No Interaction Terms, Spline Included

A critic may object to our Case 1 results on the grounds that no spline was included. However, the addition of a spline raises the number of explanatory variables to 84, thereby precluding a direct implementation of our MC^3 algorithm. Accordingly, we describe a new method for implementing Bayesian model averaging. With traditional Bayesian model averaging, the set of models is defined by whether each variable is included or excluded. If K is the number of potential explanatory variables, then we have 2^K models. For K greater than 50, the number of models is simply too large. Even using an MC^3 algorithm (and allowing the computer to run for days), does not allow for accurate estimation of the relevant posterior model probabilities. However, nothing in the theory underlying either Bayesian model averaging or the MC^3 algorithm implies that models are defined by the inclusion or exclusion of a single explanatory variable. Consequently, we can define our models in terms of whether groups of explanatory variables are included or excluded. If G is the number of groups, then we have 2^G models. In Case 2, we define our models as being dependent on whether groups of two explanatory variables are included or excluded. So even though K=84, we have G=42 and Bayesian model averaging is computationally feasible.

The format of our problem suggests a logical way to choose groups of two explanatory variables: For each explanatory variable we take the current and first lag of each of the original variables as a group and the second and third lags as another. So, for instance, different models are defined by whether O_3 does or does not have a short run effect (i.e. whether today's and yesterday's levels of ozone have explanatory power for mortality). Other models are defined by whether O_3 does or does not have a medium run effect (i.e. whether levels of ozone two and three days ago have explanatory power for mortality). Previously, we defined our models treating O_3 today, O_3 yesterday, O_3 two days ago and O_3 three days ago as separate explanatory variables; here we group the ozone lags into two separate groups. The same grouping holds for each pollutant and meteorological variables.

For the spline, we use the same strategy and define models by including or excluding groups of two knots. However, the knots are reordered so that we do not drop two knots in a row. To be specific, if the knots are originally ordered as 1,2,3,4,..,N, we re-order them as 1,3,5,..,N-1,2,4,6,..,N. So knots 1,3 are grouped together, as are knots 2,4, etc..[10]

Table 3 presents the posterior mean and standard deviation of the effect of each pollutant on mortality.

[10] In an early version of this paper, we used a two-stage strategy suggested by Clyde (2000). In the first stage, we carried out Bayesian model averaging using the spline explanatory variables and calculated the posterior mean of the trend. In the second stage, we carried out Bayesian model averaging using the air pollution and meteorological variables (and their lags) as explanatory variables with the posterior mean of the trend added as a single extra explanatory variable. A problem with this strategy is that it may over-fit the unknown trend. Nevertheless, empirical results using this two-stage approach are qualitatively similar to those presented in this paper.

Results are basically the same as in Case 1, in the posterior means are all positive, but very small relative to posterior standard deviations. The main difference between Case 1 and Case 2 is that the effects are smaller in the latter case. This is as we would expect. Case 1 omits the spline and, thus, long term and other trends in mortality could be incorrectly attributed as being due to air pollution. Thus, the health effects of air pollution in Table 2 could be overestimates.

Table 3: Cumulative Effect of Each Pollutant on Mortality

	Posterior Mean	Posterior Standard Deviation
SO_2	0.013	0.088
CO	0.004	0.045
NO	0.001	0.018
NO_2	0.027	0.123
O_3	0.020	0.102
$PM_{2.5-10}$	0.002	0.032
$PM_{2.5}$	0.017	0.098

For the sake of brevity, we do not present additional empirical results for this case. Suffice it to note that they are qualitatively similar to those for Case 1. A table comparable to Table 1 would indicate little support for any of the pollutants being in the model. The meteorological variables temperature and pressure do, however, come through strongly (as do several of the spline terms). Many models receive appreciable probability, making it is risky to choose a single model. This fact once again highlights the importance of model averaging.

3.4 Case 3: Including Interactions Terms and Spline

An important, and relatively unexplored, avenue by which air pollution may affect health is through interactions between various pollutants or between various pollutants and meteorological variables. Ideally, we should implement Bayesian model averaging using pollutants, meteorological variables, interactions and a spline. However, as discussed, such a strategy involves using 348 potential explanatory variables – and, even with MC^3 algorithms – the implementation of Bayesian model averaging (as described in the Technical Appendix) is not computationally feasible with this number of variables. In Case 3, we reduce the number of potential explanatory variables and then carry out Bayesian model averaging in a way using the grouping strategy described in

First, we begin by omitting NO, NO_2, CLOUD, HUMIDITY and WIND variables from the analysis. These are not commonly used in previous research. Our previous empirical results also found that these variables were never important. Even with these omissions, we still have five pollutants and two weather variables. Allowing for interactions between all of these variables, the inclusion of three lags and the spline, we are left with 148 potential explanatory variables; still far too many explanatory variables for implementation of traditional Bayesian model averaging. Accordingly, we use the grouping strategy outlined in Case 2 with G=4. The format of our problem suggests a logical way to choose groups of 4 explanatory variables: Take the current and three lags of each of the original variables. So, for instance, our models are defined by whether O_3 does or does not enter into the analysis in any form (i.e. either through current day's value or through any of the last three days).

Empirical results using this strategy are presented in Table 4. This table is organized such that the diagonal elements are the probability that the original explanatory variables (including the current value and three lags) should themselves be included. The off-diagonal elements refer to the interaction terms. So, for instance, the number in the cell for the row labelled "O_3" and the column labelled ”TEMP” is the probability that the interaction of ozone and temperature (including the current value and three lags) should be included. For the spline, we use the same strategy as in Case 2 (except with G=4).

Table 4: Probability of Including Each Group of Explanatory Variables

Explanatory Variable	SO_2	CO	O_3	$PM_{2.5-10}$	$PM_{2.5}$	PRESSURE	TEMP
SO_2	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CO		0.000	0.000	0.000	0.000	0.000	0.000
O_3			0.000	0.000	0.000	0.000	0.000
$PM_{2.5-10}$				0.000	0.000	0.000	0.000
$PM_{2.5}$					0.000	0.000	0.000
PRESSURE						0.807	0.000
TEMP							1.000

With the exception of temperature and barometric pressure, Table 4 is composed of zeros (to three decimal places). None of the groups of explanatory variables relating to pollutants or interactions involving pollutants has enough explanatory power for the Bayesian model averaging procedure to include them. Given our previous results, this finding is perhaps not surprising. Our previous results indicate that it is difficult enough to find an individual explanatory variable with sufficient explanatory power to warrant inclusion. Finding a group of several explanatory variables with enough joint explanatory power to warrant inclusion will be even more difficult. We do not present further results for this case since, in light of Table 4, the cumulative effect on mortality for each pollutant or interaction term will be estimated as zero.

3.5 Case 4: Restricted Sets of Explanatory Variables

One could criticize the results in the previous sections on the grounds that each involved making a compromise that made the empirical exercise computationally feasible. In Case 1, the spline and interaction terms were omitted. In Cases 2 and 3, more explanatory variables were included but compromises were made with respect to the statistical methodology used. Furthermore, one could argue that, by including so many pollutants, it is difficult to find statistical evidence in favor of the inclusion of any single pollutant. This latter criticism is probably not a valid one since the various pollutants are not that highly correlated with one another. The highest correlation (0.76) occurs between CO and NO, but most of the correlations are much lower. Nevertheless, some researchers may be interested in empirical results based on fewer pollutants.

In light of these potential criticisms, we have carried out extensive empirical work using subsets of the original explanatory variables, and our qualitative results are always similar to those above. That is, point estimates indicate that air pollution tends to increase mortality by a small amount. However, once model uncertainty is accounted for in the analysis, posterior standard deviations are very large relative to point estimates. Thus, it is impossible to conclude that air pollution has a statistically significant effect on mortality and – most importantly – it is dangerous to use point estimates as a basis for policy prescription. We do not want to place too much weight on these results based on restricted sets of explanatory variables since, by using subsets of our original explanatory variables, we are moving in the very direction that we have been so critical of earlier in this paper. That is, by discarding variables because they seemed to be unimportant in Case 1 and Case 2, we are adopting a strategy which shares some similarities with the pre-testing procedures we have criticized above. However, we present some of these results here in order to convince the reader of the robustness of our results.

For the sake of brevity, we do not present results for all of the specifications. Rather, we present results of seven Bayesian model averaging exercises where the pollutants are included one at a time. That is, we implement BMA seven different times using seven different sets of explanatory variables. Each of these contains a single pollutant along with the same set of meteorological variables and spline. Since we have seven pollutants, this leads to seven different implementations of BMA.

Thus, each set of explanatory variables contains a pollutant. Previous results indicate pressure and temperature are the most important meteorological variables and hence we include them. Including current values and three lags of these three explanatory variables (i.e. a pollutant plus temperature and pressure) along with interactions results in 24 explanatory variables. Adding a slightly coarser spline with a knot located every three months results in 24 more potential explanatory variables, or 48 in total. With this number it is possible to directly implement Bayesian model averaging in the manner described in the Technical Appendix without making any of the compromises of previous cases.

Table 5 presents the posterior means and standard deviations of the cumulative effect of each pollutant and interactions involving each pollutant. It can be seen that the magnitude of the pollution effects are (as expected) a bit larger, but are still not large relative to posterior standard deviations. That is, the posterior mean of the cumulative effect of each pollutant is never much more than one posterior standard deviation from zero (and often much less). In a similar fashion, there is little evidence for the importance of any of the interactions. Although for ozone and coarse particulate matter, there is some weak evidence (posterior mean roughly 1.5 posterior standard deviations from zero) that the interaction between the pollutant and temperature might have some affect on mortality. Although we do not report them here, it is worth mentioning that the meteorological variables have a strong explanatory role (as do many of the terms in the spline).

To provide a more intuitive interpretation of our results, let us focus on ozone. Since our explanatory variables are standardized to have mean zero and standard deviation of one, the interpretation of the cumulative effect of O_3 on mortality may be expressed as follows: If unusually high levels of O_3 are sustained for at least three days (i.e. the level of ozone is two standard deviations above its mean), then the point estimate in Table 5 suggests an increase in daily mortality of over half a death per day (i.e. $2 \times 0.268 = 0.536$). Since the average number of daily deaths in Toronto during our time period is only 47, this point estimate is fairly substantial. However, the posterior standard deviation associated with this measure is 0.598 which is very large relative to the magnitude of the effect. Thus, once again we can conclude that, when model uncertainty is taken into account, point estimates are extremely unreliable and should not alone be used for policy purposes.

Table 5: Cumulative Effect of Each Pollutant and Interactions involving a Pollutant on Mortality

	Posterior Mean	Posterior Standard Deviation
Included Pollutant is SO_2		
SO_2	0.232	0.275
$SO_2 \times$PRESSURE	0.003	0.048
$SO_2 \times$TEMP	0.015	0.123
Included Pollutant is CO		
CO	0.278	0.278
CO×PRESSURE	−0.004	0.052
CO×TEMP	0.151	0.236
Included Pollutant is NO		
NO	0.022	0.090
NO×PRESSURE	0.013	0.063
NO×TEMP	−0.023	0.184
Included Pollutant is NO_2		
NO_2	0.322	0.296
$NO_2 \times$PRESSURE	0.017	0.075
$NO_2 \times$TEMP	0.093	0.185
Included Pollutant is O_3		
O_3	0.268	0.299
$O_3 \times$PRESSURE	0.003	0.071
$O_3 \times$TEMP	0.354	0.256
Included Pollutant is $PM_{2.5-10}$		
$PM_{2.5-10}$	0.087	0.169
$PM_{2.5-10} \times$PRESSURE	0.002	0.057
$PM_{2.5-10} \times$TEMP	0.340	0.242
Included Pollutant is $PM_{2.5}$		
$PM_{2.5}$	0.295	0.308
$PM_{2.5} \times$PRESSURE	0.039	0.116
$PM_{2.5} \times$TEMP	0.176	0.233

For the case where ozone is the single pollutant used, a key message of this paper – model averaging can yield results which are substantially different from those based on a single model – is particularly relevant. Of the 2^{48} models we consider, the most probable model contains today's level of ozone as an explanatory variable (as well as several other explanatory variables). A noninformative prior Bayesian (or a non-Bayesian researcher using OLS) who presented results for this most probable model would estimate the effect of ozone on mortality to be 0.526 (much larger than that presented in Table 5) with a posterior standard deviation (or standard error for the non-Bayesian) of 0.176. For the non-Bayesian, this would translate into a very significant t-statistic of 2.987. Thus, the Bayesian or non-Bayesian researcher who used only this single regression would conclude that ozone has a sizeable and strongly significant effect on mortality. Note, however, that this most probable model receives only 0.23% of the probability. Thus, the researcher who chooses this single model would be ignoring hundreds of other, almost equally plausible models (many of which imply that ozone has no effect on mortality). In such circumstances, where a myriad of potential explanatory variables exist and there is thus great uncertainty over which model is the correct one, it is very important that empirical results incorporate this uncertainty.

3.6 Comparison with Related Work

There are so many other studies of the effect of air pollution on mortality using daily time series data that a thorough comparison of our work with others is not possible. However, the vast majority of studies have found that air pollution has a positive and statistically significant effect on mortality. For instance, Environment Canada scientists carried out a literature survey with regards to the effect of ozone on mortality and concluded:

> On balance, the time series studies examined in this analysis indicate that the association between ozone and mortality is positive, consistent and independent of other co-occurring air pollutants including particulate matter. Seventeen of the 23 studies examined reported statistically significant independent associations using single pollutant models. Fourteen studies reported results using multi-pollutant models, eleven of which demonstrated statistically signficant independent associations between ozone and mortality. (NAAQO, 1999).

Environment Canada scientists reached similar conclusions for other pollutants.

However, other recent studies, which use more sophisticated statistical methods, indicate that the findings of the present study are not unreasonable. Clyde (2000) is the work most closely related to our own. This paper uses Bayesian model averaging on a different data set (particulate matter for Birmingham, Alabama) with different explanatory variables, but finds results qualitatively similar to those in the present study. That is, point estimates indicate particulate matter has a positive effect on mortality but 95% posterior density intervals include points of no effect. Clyde (2000) presents relative risks (so that a value of 1.0 indicates a pollutant has no effect on mortality) and concludes "Relative risks based on a 10 μg/m^3 change.... lead to [95%] intervals of (0.995,1.016) under [a prior based on AIC] and (0.999,1.011) under [a prior based on BIC] using Bayesian model averaging".[11]

Another influential recent study is Dominici, Samet, and Zeger (2000) which pools data from 20 US cities using a Bayesian hierarchical modeling strategy. This paper presents a range of results for different specifica-

[11] Many researchers present the health effects of air pollution in terms of the effect of a 10 unit change (e.g. 10 μg 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 the reader interested in translating our results, for our pollutants one standard deviation is 3.54 ppb for SO_2, 0.29 ppm for CO, 21.30 ppb for NO, 8.30 ppb for NO_2, 9.15 ppb for O_3, 4.86 micrograms per m^3 for $PM_{2.5-10}$ and 8.75 micrograms per m^3 for $PM_{2.5}$.

tions, but overall find positive and statistically significant associations between pollutants and mortality. For instance, for their baseline model they conclude: "Overall, we found that a 10 μg m^{-3} increase in [particulate matter] is associated with an estimated 0.48% increase in mortality (95% interval: 0.05 , 0.92)". A point to note here is that, even though results are "positive and significant" in the sense that the point estimate is positive and the 95% interval does not include zero, the effect of particulate matter is very imprecisely estimated. Hence, even without Bayesian model averaging, this extensive study finds it hard to estimate the effect of air pollution on health precisely.

4 Conclusion and Discussion

The main objective of this paper was to carry out an empirical investigation of the effects of air pollutants on mortality using a hitherto largely unanalyzed and extensive data set and an appropriate econometric methodology that takes into account the uncertainty about precisely which explanatory variables should be included in the analysis. Our main empirical finding can be summarized thus: Point estimates of the effect of numerous air pollutants on mortality all tend to be positive, albeit small. However, when model uncertainty is accounted for in the analysis, measures of uncertainty associated with these point estimates become very large. Indeed they become so large that the hypothesis that air pollution has no effect on mortality is a plausible one. On the basis of these findings, we definitely recommend against the use of point estimates from time series studies for setting regulatory standards for air pollution exposure, at least in our Toronto case study.

A further purpose of this paper was to investigate whether interactions between different pollutants or between various pollutants and meteorological variables may determine air pollution-mortality effects. Before carrying out the empirical analysis, we argued that such interactions have been largely overlooked in the literature and are of potentially great importance. However, our empirical results indicate that these interactions are not so important, at least for the data set under consideration.

We stress that these findings do not necessarily imply that air pollution has no adverse effects on health (or the corollary, that air pollution abatement and regulatory policies should not take into account non-mortality related effects such as potential impacts on asthma and other respiratory illnesses). Rather, our results indicate that there is no reliable statistical evidence for a link between air pollution and mortality. Part of this finding may be attributed to the standard criticisms of time series studies involving daily mortality measures (i.e. the so-called "ecological fallacy" and the fact that such studies will, at best, measure only short term air pollution effects). Kunzli et al (2001) argue that time series studies will underestimate the effect of air pollution on health for these reasons, and thus, strongly recommends the use of cohort studies. Discussant comments on Dominic, Samet and Zeger (2000) provide a useful summary of these criticisms, along with a response by the authors in favor of time series studies. In defence of their position, they stress that time series studies have been used extensively to determine pollution standards for exposure. If for no other reason than for the design of air quality regulations (the implementation of which incurs immense economic costs), it is important that researchers use appropriate statistical methods to estimate air pollution impacts. Indeed, proper treatment of model uncertainty should be an essential part of any statistical method.

Furthermore, all of the models in this paper allow for all explanatory variables to enter in a linear fashion. It is possible that significant health effects only occur when air pollution levels increase beyond a threshold. If this threshold is sufficiently high, then linear models may miss important health effects. The evidence on whether thresholds exist in air pollution-mortality relationships is mixed (see Dominici, Daniels, Zeger and Samet, 2002 and Pope, 2000). Nevertheless, its is an important topic for future research. Given uncertainty about what might trigger threshold effects (e.g. is it due to the average level of a certain pollutant over several days? The interaction between two pollutants? A cumulative buildup of pollutants over many days? A single high pollution level?), the use of Bayesian model averaging is called for.

In this paper, we have presented empirical work relating to a particular environmental problem. However, a subsidiary aim of this paper has been to sell an econometric methodology to researchers working on a wide variety of problems in environmental economics. Uncertainty over which model is the appropriate one pervades many empirical applications in this field. As this paper has shown, fundamental empirical results can be sensitive to the treatment of model uncertainty. Ignoring this issue can lead the researcher seriously astray. Fortunately, Bayesian model averaging allows for a formal treatment of model uncertainty. Implementation of Bayesian model averaging, albeit more difficult than simply running a regression, is not too difficult and should be well within the abilities of applied economists. In short, this is a practical, relatively simple, econometric methodology well-suited for and immensely important for applied work.

5 References

Alberini, A., Cropper, M., Fu Tsu-Tan, Krupnick, A., Liu, Jin-Tan, Shaw, D. and Harrington, W. (1997). "Valuing health effects of air pollution in developing countries: The Case of Taiwan," *Journal of Environmental Economics and Management*, 34, 107-126.

Burnett, R., Brook, J., Yung, W., Dales, R. and Krewski, D. (1997). "Association between ozone and hospitalization for respiratory diseases in 16 Canadian cities," *Environmental Research*, 72, 24-31.

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

Clyde, M. and DeSimone-Sasinowska, H. (1997). "Accounting for model uncertainty in Poisson regression models: Particulate Matter and Mortality in Birmingham, Alabama," Institute of Statistics and Decisions Sciences, Duke University, Discussion Paper 97-06.

Clyde, M., Guttorp, P. and Sullivan, E. (2000). "Effects of ambient fine and coarse particles on mortality in Phoenix, Arizona," Institute of Statistics and Decision Sciences, Duke University*,* Discussion Paper 00-05.

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

Dasgupta, S., Laplante, B., Wang, H. and Wheeler, D. (2002). "Confronting the environmental Kuznets curve," *Journal of Economic Perspectives*, 16, 147-168.

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.

Delucchi, M.A., Murphy, J.J. and McCubbin, D.R. (2002). "The health and visibility costs of air pollution: A comparison of estimation methods," *Journal of Environmental Management*, 64, 1-39.

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

Dominici, F., Samet, J. and Zeger, S. (2000). "Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy (with discussion)," *Journal of the Royal Statistical Society, Series A*, 163, 263-302.

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

ENS (Environmental News Service). (2002). Supreme Court Clears EPA to Regulate Cleaner Air. http://ens.lycos.com/ens/feb2001/2001L-02-27-07.html

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.

Hahn, R.W. (2000). "The Impact of economics on environmental policy," *Journal of Environmental Economics and Management*, 39, 375-399.

Hoek, G., Schwartz, J., Groot, B. and Eilers, P. (1997). "Effects of ambient particulate matter and ozone on daily mortality in Rotterdam, The Netherlands," *Archives of Environmental Health*, 52, 455-463.

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.

Khanna, M., and Damon, L.A. (1999). "The EPA's Voluntary 33/50 Program: Impact on toxic releases and the economic performance of firms," *Journal of Environmental Economics and Management*, 37, 1-25.

Kunzli, N., Medina, S., Kaiser, R., Quenel, P. Horak, F. and Studnicka, M. (2001). "Assessment of deaths attributable to air pollution: Should we use risk estimates based on time series or cohort studies," *American Journal of Epidemiology*, 153, 1050-1055.

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

NAAQO (1999). "National ambient air quality objectives for ground level ozone: Science assessment document. Report by the federal-provincial working group on air quality objectives and guidelines," Environment Canada.

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

Navrud, S. (2001). "Valuing health impacts from air pollution in Europe," *Environmental and Resource Economics*, 20, 305-329.

Ostro, B., Hurley, S. and Lipsett, M. (1999). "Air pollution and daily mortality in the Coachella Valley, California: A study of PM10 Dominated by Coarse Particles," *Environmental Research, Section A*, 81, 231-238.

Poirier, D. (1995). *Intermediate Statistics and Econometrics: A Comparative Approach*, Cambridge: The MIT Press.

Pope, C. (2000). "Invited commentary: Particulate matter-mortality exposure-response relationships and thresholds," *American Journal of Epidemiology*, 152, 407-412

Ross, A., Ponce de Leon, A., Bland, J., Bower, J., Strachan, D. (1996). "Air pollution and daily mortality in London, 1987-92," *British Medical Journal*, 312, 665-669.

Sala-i-Martin, X. (1997). "I just ran two million regressions," *American Economic Review*, 87, 178-183.

Samet, J., Zeger, S., Kelsall, J., Xu, J. and Kalkstein, L. (1998). "Does weather confound or modify the association of particulate air pollution with mortality?" *Environmental Research, Section A*, 77, 9-19.

Schennach, S. (2000). "The economics of pollution permit banking in the context of Title IV of the1990 Clean Air Act amendments," *Journal of Environmental Economics and Management*, 40, 189-210.

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

6 Data Appendix

In this study, we use daily data on mortality, pollutants and meteorological variables from 1992-1997. We discuss each of these in turn. 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 circulatory 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 the data used here, 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 (results not reported in this paper since they were similar to those found using total mortality), we coded all suppressed values as the average of 1 and 4 (i.e. 2.5).

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, e.g., Delfino et al, 1994, page 22).

- Wind speed (km per hour).

There are very few missing values in the hourly data. These are replaced by a simple average of values for the hours before and after the missing value. To create daily data from this hourly data, we simply 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 which have nearly complete data for 1992-1997. These monitors are widely dispersed across Metro Toronto: In downtown Toronto (Bay/Grosvenor), Scarborough (Lawrence/Kennedy), North York (Yonge/Finch), Etobicoke (Elmcrest Rd.), Etobicoke (Evans/Arnold) and York (Clearview/Keele). Missing values are handled in the same manner as for the weather variables. There are relatively few missing values with the worst monitor having 2% of 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 being less than 2.5 micrograms while coarse particulate matter ($PM_{2.5-10}$) is defined between 2.5 and 10 micrograms. For the years 1992-1994 the only available monitor was at Bay/Wellesley. For 1996-1997 the only available monitor was at Evans/Arnold. 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 involving particulate matter since 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 of Delfino et al (1994). In particular, using the particulate matter values which are not missing, we run a regression using relevant explanatory variables. We then use the values of the explanatory variables on the missing days and 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) suggest 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 $PM_{2.5-10}$ 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 where the particulate matter variables are the dependent variables.

7 Technical Appendix

We implement Bayesian model averaging using the approach outlined in Fernandez, Ley and Steel (2001b), using the MC^3 algorithm developed in Madigan and York (1995). The reader is referred to these papers (see also Hoeting et al, 1999) for details beyond those presented in this appendix.

We have data for $t = 1, .., T$ days[12] and denote data on the dependent variable (mortality) by $y = (y_1, .., y_T)'$. All the potential explanatory variables (including lags) are stacked in a $T \times K$ matrix X. We have $r = 1, .., R$ models, denoted by M_r. These are all Normal linear regression models which differ in their explanatory variables,

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

[12] When p lags are included in the model, we proceed conditionally upon p initial observations and, hence, y_1 will actually be the p^{th} day of January, 1992.

where ι_T is a $T \times 1$ vector of ones , X_r is a $T \times k_r$ matrix containing some (or all) columns of X. The $T-$vector of errors, ε, 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.

The models are thus defined by their choice of explanatory variables (i.e. by the choice of X_r). The standard approach to Bayesian model averaging assumes different models are defined by the inclusion or exclusion of each variable. This leads to 2^K models. If K is at all large, the enormous number of potential models imposes commensurately enormous computational demands. It is worth noting that these computational demands help motivate our choice of the Normal linear regression model. Other work with daily mortality counts often uses Poisson regression methods, but this would greatly add 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 which looks Normal. The considerations suggest that the costs associated with working with a Normal model are small (i.e. our dependent variable takes on so many values and has a roughly bell-shaped histogram that its discrete distribution can be very well approximated by a continuous Normal distribution).

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:

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

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

$$g_r = \left\{ \begin{array}{l} \frac{1}{K^2} \text{ if } T \leq K^2 \\ \frac{1}{T} \text{ if } T > K^2 \end{array} \right. . \tag{A.5}$$

The empirical results in this paper use this choice for g_r, although the other choices they consider lead to qualitatively similar results.

The resulting posterior for β_r follows a multivariate t-distribution with mean:

$$E\left(\beta_r | Data, M_r\right) \equiv \overline{\beta} = \overline{Q} X_r' y, \tag{A.6}$$

covariance matrix:

$$var\left(\beta_r | Data, M_r\right) = \frac{\overline{\nu s}^2}{\overline{\nu} - 2} \overline{Q} \tag{A.7}$$

and $\overline{\nu} = N$ degrees of freedom. Furthermore,

$$\overline{Q} = \left[\left(1 + g_r\right) X_r' X_r\right]^{-1}$$

and

$$\overline{s}^2 = \frac{\frac{1}{g_r+1} y' P_{X_r} y + \frac{g_r}{g_r+1} \left(y - \overline{y}\iota_T\right)' \left(y - \overline{y}\iota_T\right)}{\overline{\nu}},$$

where

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

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

$$p\left(M_r | Data\right) = c \left(\frac{g_r}{g_r + 1}\right)^{\frac{k_r}{2}} \left[\frac{1}{g_r + 1} y' P_{X_r} y + \frac{g_r}{g_r + 1} \left(y - \overline{y}\iota_T\right)' \left(y - \overline{y}\iota_T\right)\right]^{-\frac{T-1}{2}} \tag{A.8}$$

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

Our parameters of interest measure the cumulative effect of a pollutant on mortality and these are a linear function of the regression coefficients. Hence, the previous equations are all that is required to carry out Bayesian model averaging as given in (2.1).

If the number of models, R, is relatively small (A.8) can be evaluated for every possible model and Bayesian model averaging can be implemented directly. In traditional applications of Bayesian model averaging, $R = 2^K$ (i.e. every possible explanatory variable can either be included or excluded). For cases where

$K > 20$ direct implementation of Bayesian model averaging is computationally infeasible. Accordingly, we adopt the MC^3 described in Madigan and York (1995). This is a Metropolis algorithm which is very simple to implement. In particular, if the current model in the chain is M_s then a candidate model, M_j, which is randomly (with equal probability) selected from the set of models including M_s and all models containing one more or one less explanatory variable (i.e. the algorithm randomly either adds or subtract one column from X_s), is drawn. M_j is accepted with probability:

$$\min\left\{1, \frac{p\left(M_j|Data\right)}{p\left(M_s|Data\right)}\right\}.$$

If M_j is not accepted then the chain stays with M_s. It can be shown that the relative frequency that each model is drawn will converge to its posterior model probability.

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.8). Then we approximate this probability using output from the MC^3 algorithm. When these probabilities are the same to three decimal places, we deem convergence to have taken place. The number of draws required for the various models considered varied from 1,000,000 to 2,000,000.