

Bayesian statistical modeling of the relationship between air quality and mortality:  
In pursuit of accurate uncertainty bands and better environmental policy

Project proposal submitted to the *University Research Program at Ford Motor Company*

David Draper (*University of California, Santa Cruz*)  
and Chris Gearhart (*Ford Motor Company*)

27 September 2005

**Introduction.** A key input to the 1997 National Ambient Air Quality Standards for particulate matter and ozone concentration was a set of epidemiological studies showing a correlation between daily mortality and short-term concentration of particulate matter. Since 1997 a number of similar studies have been performed for different regions of the country using different modeling schemes and different measures of particulate matter concentration. Unfortunately, as Chock, Winkler and Chen (2000: *Journal of the Air and Waste Management Association*, 50, 1481–1500) note:

“Depending on how one deals with the choice of weather variables, co-pollutants, smoothing schemes, seasonal considerations, and so forth, it is possible and in fact not surprising that different best-effort models applied to data sets from the same urban areas can reach different or even contradictory conclusions. ... There are at least two reasons for the confusion [arising from this lack of consensus]. First, the sought-after signal is very weak compared to the noise in the data, so that model results become sensitive to the choice of the model. ... Second, there are serious confounding and multicollinearity effects due to the presence of correlations among co-pollutants, for example, and other potentially causal variables that may or may not be included in the models.”

Our concern in proposing this research is that future regulatory requirements may be based on a small number of epidemiological studies with a limited number of co-predictors included. Such studies will likely underestimate the uncertainty in their conclusions and therefore overestimate the risk to society. In principle erring on the conservative side generally is viewed as a good idea when it comes to public health, but this oversimplifies the situation. All organizations have to work with a limited set of resources. Engineering effort required to meet very strict emissions requirements that may have little benefit to society will inevitably take valuable resources away from safety and fuel economy improvement efforts that have a much larger societal benefit.

The goal of this study is to establish good methods for estimating the true uncertainty in epidemiological studies of the relationship between ambient air pollutant levels and mortality, which will serve as the basis of better environmental policy.

**Methods.** Our basic idea is to take advantage of recent advances in computing speed and Bayesian statistical analysis to develop a more comprehensive modeling framework that comes closer to quantifying *all* relevant sources of uncertainty and propagating the totality of uncertainty through to the best possible decision. As Chock, Winkler and Chen (2000) mentioned in the quote above, when attempting to draw valid causal conclusions from observational data in the presence of multiple confounding, “model results become sensitive to the choice of the model,” and different investigators can draw different conclusions from the same data. This is because each set of investigators, using traditional statistical modeling tools such as standard regression and time series methods, fails to properly account for the *uncertainty in the model itself*: both its structure (e.g., the precise class of time series or regression models employed) and the set of co-predictors and pollutants included. *Bayesian model averaging* (e.g., Draper 1995: *Journal of the Royal Statistical Society, Series B*, 57, 45–97) is a relatively recent technique that solves this problem by (1) simultaneously embracing many models for the same data, (2) allowing each model to produce its own predictive distribution, and then (3) creating a composite predictive uncertainty band by taking a weighted average of each

