APPENDIX AVAILABLE ON REQUEST

Special Report 17

Traffic-Related Air Pollution: A Critical Review of the Literature on Emissions, Exposure, and Health Effects

Chapter 3. Assessment of Exposure to Traffic-Related Air Pollution

HEI Panel on the Health Effects of Traffic-Related Air Pollution

Appendix C. Bayesian Hierarchical Modeling

Correspondence may be addressed to Dr. Maria Costantini, Health Effects Institute, 101 Federal Street, Suite 500, Boston, MA 02110, mcostantini@healtheffects.org.

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83234701 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and therefore may not necessarily reflect the views of the Agency, and no official endorsement by it should be inferred. The contents of this document also have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties, and no endorsement by them should be inferred.

This document did not undergo the HEI scientific editing and production process.

APPENDIX C. BAYESIAN HIERARCHICAL MODELING

The framework of Bayesian hierarchical modeling (BHM) refers to a generic model-building strategy in which unobserved quantities (e.g., statistical parameters, missing or mismeasured data, random effects, etc.) are organized into a small number of discrete levels with logically distinct and scientifically interpretable functions and probabilistic relationships between them that capture inherent features of the data. It has proved to be successful for analyzing many types of complex epidemiologic, biomedical, environmental, and other data, as illustrated by the varied case studies in Gilks et al. (1996) and Green et al. (2003). When specifying a hierarchical model, it is often convenient to start with a graphic representation of the structural assumptions that relate to the quantities in the model. Such models are commonly referred to as "Bayesian graphical models" and have become increasingly popular as "building blocks" for construction of complex statistical models (Spiegelhalter 1998). In the study of the health effects of air pollution, Bayesian hierarchical models have been used for a variety of purposes, foremost by Dominici and co-investigators in a series of papers, starting with Dominici, et al. (2000), that present combined analyses of time-series data in U.S. cities and use the BHM framework to summarize the information. These models also have been used in the context of exposure models, for example, to carry out a synthesis of pollution measurements and model-based estimates (Fuentes and Raftery 2005) or to model jointly multivariate pollutants at different sites (Shaddick and Wakefield 2002). The general applicability of BHM has been enhanced by advances in computational algorithms, notably those belonging to the family of stochastic algorithms based on Markov chain Monte Carlo (MCMC) techniques that have enabled the estimation of custom-specified models (for a review of recent algorithmic developments, see Green et al 2003).

We note that BHM encompasses structures that are generally referred to as multilevel models, where the data have a nested structure, for example, subject, school, community, as well as a time indexing. Multilevel structures typically are specified with random effects models, and those — as well as much more general structures — can be accommodated within the BHM framework with the use of efficient MCMC algorithms to simulate the full posterior distributions of all unknown parameters (Gelman et al 2003), without recourse to approximation and without imposing restricted assumptions like Gaussian distributions for the random effects. A key difference between Bayesian and likelihood-based mixed-model estimation of multilevel data is the treatment of uncertainty of key quantities such as random effects variances.

The main quantities that are involved in modeling relationships between environmental exposures to air pollution (e.g., those related to traffic) and health are the health outcome, Y, the true exposure, X, and the surrogate measure(s), Z. The quantities Y, X and Z can be measured either at the individual level or associated with a group — for example, a geographical unit. Broadly speaking, statistical modeling is used to specify a "disease model" that links Y and a latent X and a "measurement model" that links X and Z. In a Bayesian framework, a prior model of the variability of the latent X needs also to be specified. Three main designs, which we refer to as *individual*, *semi-ecological,* and *ecological* (or aggregate) designs, have been employed in environmental epidemiology dependent on the type and resolution of the available data. The analysis of each of these designs can gain from several key features of BHM that make this model-building strategy attractive in the field of environmental health.

In general terms, the benefits of BHM can be broken down under a number of headings, with obvious overlap between some of these:

(i) Modular model elaboration
(ii) Integration of different sources of information
(iii) Coherent propagation of uncertainty
(iv) Borrowing of strength
(v) Integrated treatment of information at different levels.

We will refer to these headings in the discussion of study designs and of some of the key issues faced by the statistical modeler in the air-pollution and health field.

The individual design, the simplest, is appropriate if both determinants of exposure and health outcome data are measured at the individual level on the same set of subjects. Apart from clinical studies of a limited number of voluntary subjects exposed to a controlled air-pollution environment, this design is not commonly used but can constitute part of the data, e.g., a validation sample, collected in an epidemiologic study. Standard application of Bayesian hierarchical modeling can be used to propagate uncertainty in the measurement of X, by linking one or several individual surrogate measures to the unknown individual exposure, X, treated as an unmeasured random variable (Richardson and Gilks 1993). An interesting variant of this design is the case where one is interested in combining individual level data with different levels of covariate details, for example, routinely collected national register data containing limited information on confounders and a detailed cohort study. Bayesian graphical models can be used to synthesize data sets with different sets of covariates to improve inference. In a recent study, Jackson et al. (2008b) evaluated the link between NO_2 and low birth weight based on a combination of register and cohort data in the U.K.

Semi-ecological designs arise when health outcome data are measured at the individual level, but measurements of exposure involve some form of group-level data (for example, time and spatially averaged ambient pollution concentrations, or modeled exposure to traffic). This design has also been referred to as semi-individual (Künzli and Tager 1997). Semi-ecological designs are commonly used, in particular in cohort studies that investigate the link between air pollution and disease risk. In this case, hierarchical modeling is useful to elaborate a suitable — often complex — measurement model that links different sources of exposure data measured, both at the individual and the ecological level. Individual level determinants could consist, for example, of geo-coded housing location, residential history, time-activity diaries, indoor measurements, or records from personal badges. To accommodate such diverse sources of individual-level exposure data as well as group-level environmental data from monitoring stations and/or modeled traffic data, a series of submodels or "modules" that link all these sources of data with the latent variable of interest, e.g., the personal exposure to traffic, have to be built. By using BHM techniques where each new type of data is treated in a modular fashion to form part of a comprehensive exposure, the analyst ensures that all relevant data will contribute to the assessment of "true"

Reproduced from Molitor et al. 2006, with permission of Oxford University Press.

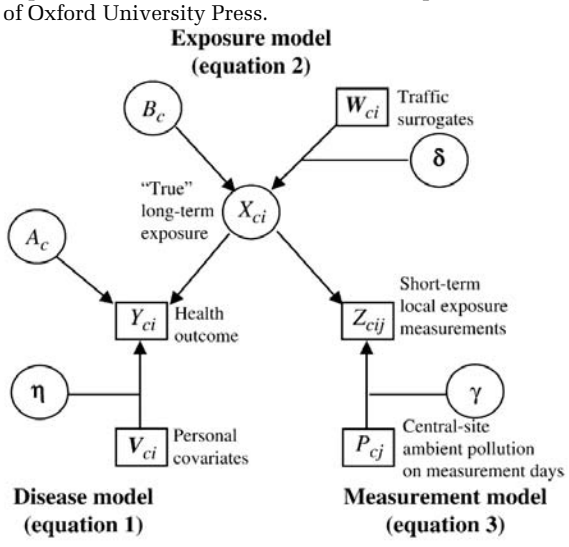

FIGURE 1. Schematic representation of the overall model used to analyze data in the Children's Health Study pilot project, conducted in the year 2000. Note that, while not represented in the graph, the random intercepts A_c and B_c are modeled by use of central-site exposure measurements P_{c*}.

exposure in a flexible fashion. Such a comprehensive exposure model typically will be a combination of Berkson and classical error structures that would be difficult to analyze with standard non–Bayesian techniques. A recent example of the power of using such a combination of measurement error models treated in a fully Bayesian way is the study of Molitor et al. (2006) in which the long-term effect of NO_2 exposure on lung function of children was analyzed. The graph reproduced from Molitor et al. (2006) shows the complex measurement error structure, which is partly Berkson (link between X_{ci} and W_{ci}) and partly classical (link between Z_{cij}, P_{cj} and X_{ci}).

Another framework that is useful for semi-ecological designs is that of point process models. In environmental epidemiology, one might consider that the location of the individuals at risk is being modeled by a baseline demographic process, on which is super-imposed a disease process that "picks"out the cases with a probability that is dependent both on individual-level risk factors and area-level variables measuring exposure (Richardson and Best 2003). In a study of the effect of traffic pollution on respiratory disorders of children carried out by Best et al. (2000), a point process is incorporated in a hierarchical model, to account both for individual risk factors of the children and for environmental levels of air pollution.

Ecological designs are appropriate when both the health outcome data and exposure data are only available at the group level. Most time-series studies use this design with the group being composed of inhabitants of a city or a small geographic area. Daily or weekly disease counts are then linked by standard Poisson regression to averaged air pollution recorded during the same time intervals or shifted by short time lags. It is well known that the shape of the exposure-effect relationship is different at the group level from that at the individual level, a phenomenon known as *ecological bias* and much debated (Greenland 1992; Richardson and Montfort 2000; Richardson et al 1987); despite this, most ecological designs rely on empirically specified relationships between aggregated counts and mean group exposure, which are difficult to interpret. Alternatively, it is possible in many cases to work out how to integrate individual-level exposure effect relationships to the group level and use these integrated forms for specification of relationships in ecological models (Richardson and Best 2003). This approach retains the interpretability of the individual-level coefficients, independent of the scale of aggregation. Building on this, hierarchical models can be used to go further and ensure a better control of ecological bias by construction of a joint analysis of individual- and group-level data, linked by shared individual-level coefficients (Jackson et al 2008a).

The modularity discussed previously renders possible the simultaneous integration of different sources of information at the personal and group level. Moreover, by building a *joint model* of all the relevant variables that can include any number of different modules, all sources of information are integrated coherently to contribute to the estimation of the exposure-effect relationships of interest, thus maximizing the use of all relevant information. In parallel, the joint hierarchical model leads automatically to a correct *propagation of all sources of uncertainty* captured in each "module" onto the estimation of the health effect parameters of interest. We stress that this is not the case for methods that proceed by substitution and plug-in, for instance by replacing an unknown individual exposure X by an estimated one, based on a regression-calibration approach. In the BHM framework, unknown exposures are treated as random variables, their distribution is informed by the different substudies to which they are related, and the associated uncertainty is then propagated on the distribution of the health-related regression coefficients of interest. It is important to note that *both* the surrogate measures Z *and* the disease information Y will contribute to posterior distribution of X and that this dual link is key to insure that the measurement error correction leads to unbiased estimators of the health effects and a realistic assessment of uncertainty. A recent study by Gryparis

et al. (2008) exemplifies this and shows that averaging over Monte Carlo simulations of the variability of X based *only* on surrogate information will lead to biased estimates of effects.

While the general principles of Bayesian measurement error modeling are certainly relevant here, the specificity of the measurement of air pollution adds additional interesting features, in particular with regards to the need to account for the spatial structure of the pollution field. A typical situation is that of spatial misalignment, i.e., a need to allocate exposure measures to geo-coded residential locations where no monitoring data are available. To construct exposure assessment, a spatial model of the pollution field can be built with data from monitoring networks in a larger area that encompasses the locations of interest and, subsequently, used to predict relevant air pollutants at locations of interest. Plugging in smoothed predicted values in a cohort-based analysis is commonly done (Hoek et al 2001) and corresponds in effect to a regression-calibration approach. While corrections can be built for compensation of the oversmoothing of the predictive values and the underestimation of the variability of the health effects estimates given by the regression-calibration approach, a full Bayesian treatment requires carrying out spatial modeling of the pollution field and disease model estimation *jointly*. By integrating the posterior distribution of pollution prediction at the required locations within the disease model, the uncertainty of exposure assessment and the spatial correlation of the errors are, thus, fully propagated onto the health effect estimate. Such a joint model implementation can be computationally demanding for a large area, and recent work has investigated alternative two-stage Bayesian approaches (Gryparis et al 2008).

So far, we have discussed how to propagate exposure measurement uncertainty in a coherent manner. However, propagation of uncertainty has to be considered at every level of the analysis; and principled Bayesian approaches should include a component of uncertainty linked to model choice. Model choice issues and how to account for these by use of Bayesian model averaging (BMA) have been discussed extensively in the statistical literature (Draper 1995; Hoeting et al 1999) and, in particular, in epidemiology (Viallefont et al 2001). Basically, posterior model probabilities are used to reweight the estimate of the quantities of interest. For time-series air-pollution studies, a specific aspect of this problem concerns the uncertainty of adjustment for unobserved time varying confounders, an adjustment which is typically carried out by incorporation of functions of calendar time and temperature with varied degree of smoothing (Peng et al 2006). The degree of smoothing is clearly influential on the effect estimate, and there is currently a lively debate on how to choose a suitable degree of smoothing. In this context, a blind application of BMA that treats symmetrically the exposure and confounders is not appropriate, since models without confounder adjustment are not epidemiologically interpretable. Simple modification of BMA that forces a set of confounders in the regression models can be used, if the confounders are known (Raftery and Richardson 1995). In other cases, suitable modifications of BMA have been proposed recently that account for adjustment uncertainty in effect estimation (Crainiceanu et al 2008), and this is a promising area for future research.

One important aspect of hierarchical models that has been highlighted so far is the integration of different sources of data. A related and complementary aspect is the *borrowing of strength* between different but related analyses of similar data sets that leads to improved and more stable estimates of the parameters of interest. We refer here to the use of Bayesian hierarchical models to perform a meta-analysis of a group of studies that are *comparable with respect to the measurement of exposure and health* outcomes. The key principles behind the use of BMH for meta-analysis have been discussed in broad terms by Sutton and Abrams (2001) and used abundantly in air-pollution epidemiology by Dominici and colleagues. BHM attempts to go beyond a standard random effects model by allowing

more flexibility of specification, the inclusion of prior information on sources of heterogeneity, and the computation of probabilistic statements on key quantities. As illustrated by Dominici et al. (2000) in their combined analysis of NMMAPS data, the crucial step in meta-analysis is the specification of a between-study model of variability, which, in turn, allows borrowing strength between the studies. Given first level estimates of effects $\hat{\beta}_i$ that are derived in each study *i*, the second level model, represented here as $p(\hat{\beta}_i \mid \eta)$, makes assumptions on the mean, variability, and distributional form p ( ) which govern the $\hat{\beta}_i$ in terms of a number of hyperparameters η (parameters of the prior distribution). These in turn will be given a prior distribution.

The simplest assumption for borrowing strength is the exchangeable model, which hypothesizes that all the parameters of interest, for example, the coefficients that quantify the association of PM_{10} on mortality, come from a common distribution, usually Gaussian: $p(\hat{\beta}_i \mid \eta) = N(\mu, \sigma^2)$, independently for all *i*, (1). Interest is centered on the global mean μ and the estimation of the variance parameter σ^2 that quantified the heterogeneity of the effects. In the Bayesian version, this variance (or its inverse, the precision) is itself given a prior distribution at a higher level of the hierarchy. At this point, two related and important aspects of model specification need to be scrutinized carefully, namely, the investigation of whether the exchangeability assumption is reasonable and, given this, whether the choice of the functional parametric form p( ) at the second level and whether the prior distribution for the hyperparameters η could be unduly influential.

The straightforward exchangeability assumption is questionable in many cases, although often it is used as a baseline model (see Dominici et al 2006). A common modification is to account for known sources of heterogeneity in the specification of the mean of the combined model, i.e., replace μ above by a regression equation $\sum \alpha_k Z_k$ where Z_k are known study specific covariates. For example, in an examination of the effect of PM_{10} on mortality in 20 of the largest cities, Dominici et al. (2000) adjust for city-specific covariates Z_k, such as the percentage of people in poverty, the percentage of people older than 65 years, and the average daily level of other pollutants. In their recent analysis of hospital admissions for cardiovascular and respiratory diseases in 108 counties, Peng et al. (2008) include the percentage of population living in an urban area in the second stage model to investigate potential effect modification by the chemical composition of $PM_{10-2.5}$. Beyond these elaborations of the mean model, other sources of heterogeneity related to important qualitative characteristics of the group of studies also can be accounted for by imposing some restriction on exchangeability. In other words, full exchangeability is replaced by an assumption of partial exchangeability within known subgroups, indexing the distribution p( ) and the hyperparameters η by the subgroup to which study *i* belong and replacing equation (1) by: $p(\hat{\beta}_i \mid \eta_g) = N(\mu_g, \sigma_g^2)$ independently for all *i* in group *g*. Regions have been often used to define such groupings, like eastern and western regions in the U.S. Of course, this presupposes that relevant subgroups have been identified based on a priori epidemiologic and environmental knowledge. Between full exchangeability of the first level estimates $\hat{\beta}_i$ and partial exchangeability of $\hat{\beta}_i$ for prespecified subgroups of studies, an interesting alternative is to model the collection of $\{\hat{\beta}_i\}$ as a joint multivariate distribution MVN(M ,Σ). This extends the simple mode (1) by allowing dependence between the $\hat{\beta}_i$, thus replacing σ^2 by a variance-covariance matrix Σ. A spatially structured Σ comes naturally to mind as a good candidate, when the meta-analysis concerns different areas or cities, and dependence between the city-specific estimates might be created by long-range components of pollution; such a spatial specification was used by Dominici et al. (2000).

Sensitivity to the functional assumptions of the second level model has been less discussed but is nevertheless important. Indeed, when there is substantial heterogeneity in the first level estimates and when these estimates have large uncertainty, one can suspect that there is interplay between the

specification of the between-study distribution and the posterior estimates of the overall effect and its variability. Instead of assuming necessarily a Gaussian distribution as in equation (1), it would be important to include in a sensitivity analysis alternatives such as heavier tail distributions, e.g., the *t*-distribution, mixture of Gaussians (cf. Richardson in the discussion of Dominici et al. 2000) or a flexible semi-parametric model of heterogeneity using Dirichlet process, as advocated by Ohlssen et al. (2007) for institutional comparisons. There is much work to be pursued in exploring and modeling heterogeneity, uncovering latent group structure, and building suitable dependent structures when combining information across studies.

In summary, in this brief and necessarily selective review, we have outlined the key benefits of using Bayesian hierarchical models in air-pollution and health studies, tried to highlight important issues of sensitivity, and pointed out areas for further research.

REFERENCES

Best NG, Ickstadt K and Wolpert RL. 2000. Spatial Poisson regression for health and exposure data measured at disparate resolutions. J Am Stat Assoc 95:1076-1088.

Crainiceanu CM, Dominici F and Parmigiani G. 2008. Adjustment uncertainty in effect estimation. Biometrika 95:635-651.

Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger SL and Samet JM. 2006. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. JAMA 295:1127-1134.

Dominici F, Samet JM and Zeger SL. 2000. Combining evidence on air pollution and daily mortality from the 20 largest US cities: A hierarchical modelling strategy. Journal of the Royal Statistical Society Series A-Statistics in Society 163:263-284.

Draper D. 1995. Assessment and propagation of model uncertainty. Journal of the Royal Statistical Society Series B-Methodological 57:45-97.

Fuentes M and Raftery AE. 2005. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. Biometrics 61:36-45.

Gelman A, Carlin JB, Stern HS and Rubin DB. 2003. Bayesian Data Analysis. Chapman and Hall, London, UK.

Gilks WR, Richardson S and Spiegelhalter DJ, eds. 1996. Markov chain Monte Carlo in Practice. Chapman and Hall, London, UK.

Green PJ, Hjort N and Richardson S, eds. 2003. Highly Structured Stochastic Systems. Oxford University Press, Oxford, UK.

Greenland S. 1992. Divergent biases in ecologic and individual-level studies. Stat Med 11:1209-1223.

Gryparis A, Paciorek CJ, Zeka A, Schwartz J and Coull BA. 2008. Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics.

Hoek G, Fischer P, nan den Brandt P, Goldbohm S and Brunekreef B. 2001. Estimation of long-term average exposure to outdoor air pollution for a cohort study on mortality. J Expo Anal Environ Epidemiol 11:459-469.

Hoeting JA, Madigan D, Raftery AE and Volinsky CT. 1999. Bayesian model averaging: A tutorial. Statistical Science 14:382-417.

Jackson CH, Best NG and Richardson S. 2008a. Bayesian graphical models for regression on multiple datasets with different variables. To appear in Biostatistics.

Jackson CH, Best NG and Richardson S. 2008b. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. Journal of the Royal Statistical Society Series A-Statistics in Society 171:159-178.

Künzli N and Tager IB. 1997. The semi-individual study in air pollution epidemiology: A valid design as compared to ecologic studies. Environ Health Perspect 105:1078-1083.

Molitor J, Molitor NT, Jerrett M, McConnell R, Gauderman J, Berhane K and Thomas D. 2006. Bayesian modelling of air pollution health effects with missing exposure data. Am J Epidemiol 164:69-76.

Ohlssen DI, Sharples LD and Spiegelhalter DJ. 2007. Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. Stat Med 26:2088-2112.

Peng R, Dominici F and Louis T. 2006. Model choice in time series studies of air pollution and mortality. Journal of the Royal Statistical Society Series A-Statistics in Society 169:179-203.

Peng RD, Chang HH, Bell ML, McDermott A, Zeger SL, Samet JM and Dominici F. 2008. Coarse particulate matter air pollution and hospital admissions for cardiovascular and respiratory diseases among Medicare patients. JAMA 299:2172-2179.

Raftery AE and Richardson S. 1995. Model selection for generalized linear models via GLIB: Application to Nutrition and Breast Cancer. In: Bayesian Biostatistics. (Berry DA and Strangl DK, eds.), pp. 321-353. M. Dekker, New York, NY.

Richardson S and Best N. 2003. Bayesian hierarchical models in ecological studies of health-environment effects. EnvironMetrics 14:129-147.

Richardson S and Gilks WR. 1993. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. Am J Epidemiol 138:430-442.

Richardson S and Montfort C. 2000. Ecological correlation studies. In: Spatial Epidemiology: Methods and Applications. (Elliott P, Wakefield J, Best N and Briggs D, eds.), pp. 111-120. Oxford University, Oxford, UK.

Richardson S, Stücker I and Hémon D. 1987. Comparison of relative risks obtained in ecological and individual studies: Some methodological considerations. Int J Epidemiol 16:111-120.

Shaddick G and Wakefield J. 2002. Modelling daily multivariate pollutant data at multiple sites. Journal of the Royal Statistical Society Series C-Applied Statistics 51:351-372.

Spiegelhalter DJ. 1998. Bayesian graphical modelling: A case-study in monitoring health outcomes. Journal of the Royal Statistical Society Series C-Applied Statistics 47:115-133.

Sutton AJ and Abrams KR. 2001. Bayesian methods in meta-analysis and evidence synthesis. Stat Methods Med Res 10:277-303.

Viallefont V, Raftery AE and Richardson S. 2001. Variable selection and Bayesian model averaging in case-control studies. Stat Med 20:3215-3230.