Bayesian Analysis (2009) **4**, Number 1, pp. 55–62

Rejoinder

Peter F. Craigmile*, Catherine A. Calder†, Hongfei Li‡, Rajib Paul§, and Noel Cressie¶

We would like to thank Christopher David Barr, Francesca Dominici, David B. Dunson, and Alexandra M. Schmidt for taking the time to discuss our article. We have grouped our responses under the various topics that were raised.

1 Data, data management, and software

We agree with Schmidt that it is essential that researchers from many diverse areas have access to affordable, but still trustworthy, software. In this research project, substantial effort went into preparing datasets. Much of the data came from different government agencies, with databases arranged in multiple formats, often including variables that were not immediately relevant to our scientific pursuits. In our work on this project, the use of SAS was essential to producing clean datasets.

A further issue that complicates data formation is that government agencies tend to arrange their data in a format fit for a single purpose, usually monitoring compliance in our case. In earlier work (Cressie et al. 2007), we investigated including global water information into our study of Environmental Protection Agency (EPA) Region 5 (consisting of six midwestern states). An example of the problems we encountered was that we could not find recorded arsenic concentrations for public water systems throughout EPA Region 5; while we were able to obtain data for Ohio, for the remainder of the states we only had the dates for which public water systems were out of compliance.

1.1 Information on time in the NHEXAS dataset

As Schmidt mentions, the temporal dimension of exposure-related measurements is often key to understanding dose-response relationships. With the exception of a few individuals who were monitored over multiple time periods, all NHEXAS measurements for an individual were confined to a seven-day period, and the collection of samples occurred at fixed time points within this sampling window. This aspect of the study design precluded us from exploring the strength of the media/exposure associations at different temporal lags and over different periods of integration. We note that despite these inherent limitations of our primary data source, substantial day-to-day variation

*Department of Statistics, The Ohio State University, Columbus, OH, USA, http://www.stat.osu.edu/~pfc/

†Department of Statistics, The Ohio State University, Columbus, OH, USA, http://www.stat.osu.edu/~calder/

‡IBM T. J. Watson Research Center, Yorktown Heights, NY, USA, http://www.stat.osu.edu/~hongfei/

§Department of Statistics, Western Michigan University, Kalamazoo, MI, USA

¶Department of Statistics, The Ohio State University, Columbus, OH, USA, http://www.stat.osu.edu/~ncressie/

 DOI:10.1214/09-BA401REJ

in an individual's exposure to metals is not expected (O'Rourke et al. 1999). That is, we expect more variation in arsenic exposure across individuals than across time for a particular individual. In addition, the urine biomarker of exposure is thought to be particularly sensitive to recent exposures (ATSDR 2005). Thus, for the NHEXAS participants, this measurement was likely responding to exposure that occurred during the seven-day sampling window.

2 Modeling issues

2.1 Incorporation of scientific knowledge into the models

In response to Schmidt's inquiry, we are glad to discuss how the scientific knowledge used in our model-building process was obtained. The statistical analyses described in our article constitute one part of a long-term collaborative research project involving statisticians at The Ohio State University and Battelle, an international science and technology enterprise headquartered in Columbus, OH. Through their work with the U.S. EPA, our Battelle collaborators have substantial experience in both the collection and statistical analysis of environmental-exposure data. In addition, their access to Battelle's world-class exposure scientists was a great asset to our project. In order to draw on Battelle's expertise in our research project's application area, we found that it was essential for our entire research team (consisting of four Ohio State faculty members, four Battelle research statisticians, and several graduate students) to meet about once per month. During these large group meetings, the various subgroups responsible for different aspects of the project reported on their progress and obtained feedback from other group members not directly involved in that particular task; much of this feedback led to improved scientific relevance of our research. Detailed minutes of each meeting were taken and archived for easy access by all team members. Our research group also held an informal 'journal club' in which we discussed relevant papers in the statistics and exposure-science literatures.

In addition to drawing on our Battelle collaborators' experience, we obtained a substantial amount of scientific knowledge used in our model-development process from various government agencies' websites, including EPA's. When we needed information that was not available on the web, we found that listed agency contacts were responsive to our inquiries and often went out of their way to assist us.

Finally, it is worth mentioning that even with all of the assistance we received in acquiring scientific knowledge in our area of application, we still had to invest a great deal of time to learn about arsenic exposure. This investment enabled us to develop statistical models that appropriately reflect the substantive knowledge in the field and to address the limitations of available data sources.

2.2 Prior specifications

Dunson and Schmidt bring up the topic of prior choice, albeit from different angles. Dunson highlights the problem of using high variance priors that leads to "almost" improper posterior distributions, and hence poor mixing and convergence of the MCMC algorithm. On the other hand, Schmidt worries that some of the priors used (such as those used for α^W and the measurement-error variances) are too informative. In terms of priors for variances, we did experiment with the choice of parameters in our inverse-gamma specifications, but except for priors that appear in the soil model, we did not consider distributions other than the inverse gamma. We found that our results seemed to be insensitive to these alternative prior specifications. We thank Dunson for highlighting very recent papers that discuss alternative prior distributions for variance parameters. We agree that these "weakly informative priors" seem promising, although we add that if we (as a community) are to move away from the mainstay of the normal-inverse gamma prior combination we need to carefully understand the alternatives before we start to use them. For example, while many of these types of priors can be made conjugate by parameter-expansion Gibbs steps, this computational tool could come at a cost of slower mixing.

In terms of the decision of fixed measurement-error variances for the LEB model, our choice mirrored the common problem in the literature that there is not enough information in the data to separate learning about the process and measurement-error variances (e.g., Carroll et al. 2006, p.184). Basing the measurement-error variances on information about the measurement errors gleaned from the NHEXAS documentation on data collection was our way to separate these sources of variability.

2.3 Handling different spatial scales and the proper CAR prior

The CAR prior is used to model the latent stream-sediment process, defined over watersheds. This spatial model is used in a purely descriptive capacity as part of our attempt to incorporate more complete soil information available from USGS's National Geochemical Survey. The model does not account for the different watershed sizes and uses one parameter to capture the idea that the latent process is smoother than if the errors were independent. Schmidt's suggestion of overlaying a regular grid and modeling a latent spatial process at the grid nodes would be one way to address the change-of-support problem, but at the expense of increasing considerably the number of latent variables to be predicted. A (possibly nonstationary) geostatistical model accounts for the change of support directly, but the covariance matrix requires inversion. Clearly, a proper calibration of the three approaches is needed, from which a comparison could be made.

2.4 Use of sampling weights

A description of the NHEXAS sampling design can be found in Whitmore et al. (1999) and Robertson et al. (1999). The use of sampling weights in a Bayesian analysis is

described in Gelman et al. (2004), Ch. 7. From the details given by Robertson et al. (1999) on the NHEXAS sampling design for Arizona, it appeared reasonable to assume an ignorable design. Consequently, the sampling weights do not enter into the analysis.

2.5 Further discussion of the global models

Barr and Dominici wish there had been "more room to discuss the global models as a group." In an earlier draft of this work, we did discuss the effect of adding the global models one at a time to the LEB model in more detail (also see Section 4.3 below). While this exercise did not fully allow us to investigate the global models as a whole, it certainly provided insight into the effect of each model upon the relationships we observed between the latent NHEXAS processes defined in the LEB model (which is of primary interest in our study). In terms of studying the global models independently, note that while the global soil model can be interpreted fully when disconnected from the LEB model (i.e., when we break the connection between the latent NHEXAS soil process and the latent global topsoil process), there is no way to interpret the water model alone since it is undefined without conditioning on the latent NHEXAS water process.

3 Model fitting

3.1 Number of parameters and MCMC convergence checking

Our model is defined in terms of many latent processes which, if counted as parameters, implies that our model has thousands of parameters. Of course, as Dunson acknowledges, many of these parameters are highly correlated, so the effective number of parameters is considerably less. In our analysis, we focused on assessing convergence of parameters that appear in lower levels of the hierarchy. For example, in the LEB model we assessed convergence of the parameters $\boldsymbol{\mu}^M$, $\boldsymbol{\beta}^M$, and $\boldsymbol{\tau}^M$ using trace plots, based on running many chains (usually 2 or 3 chains, but often many more as we studied the parameter nonidentifiability issues discussed in Section 4.3 of our article). For some of the global components of the model (especially soil), we did use the Gelman-Rubin (Gelman and Rubin 1992; Brooks and Gelman 1998) convergence statistics to aid in verifying convergence of the chains. We always confirmed that there was no evidence of a lack of convergence before making inferences on any model parameters.

3.2 MCMC sampling schemes

Both Schmidt and Dunson discuss the idea of using marginalization to improve mixing and, consequently, convergence of the MCMC schemes; we agree with this strategy generally. In our models, the complication with marginalizing was related to the abundance of data observed below the minimum detection limit. Although it is possible to marginalize, we chose not to follow this strategy. Further research would also be required to see if marginalization would help in the global models.

4 Results

4.1 Relationship to prior work by Clayton et al.

All three discussions comment on the fact that our analysis relies heavily on Clayton et al. (2002)'s exposure-pathways structure, which is the basis for the acyclic directed graph (ADG) underlying the LEB model. Our decision to use this structure was driven by the desire to incorporate scientific knowledge external to the NHEXAS data into our analysis. Given that Clayton *et al.* describe their pathways structure as being "generic" and suggest that it be considered as a "template" for multimedia exposure studies, we feel that our use of their structure was a valid mechanism for drawing upon their expertise – the links between media (shown in our Figure 3) were chosen by exposure experts based on what they felt were logical causal relationships between variables. These links were confirmed by our Battelle collaborators.

We did not use NHEXAS data to develop our process model. Instead, our exploratory analyses (e.g., Figure 4) were designed to be preliminary assessments of the extent to which the NHEXAS data captured the direct relationships between media proposed by Clayton *et al.*; we did not modify the underlying graphical structure (process model) based on our exploratory analyses despite apparent conflicts between the data analysis and model structure. As an example of such a conflict, Barr and Dominici note that that even though the pairwise correlation between arsenic levels in indoor air and food casts doubt on the hypothesized direct relationship between the two media, we chose to keep the indoor-air-to-food link in the LEB model. In doing so, we were able ultimately to assess the strength of the direct linear relationship between indoor air and food after adjusting for the effects of other media, in addition to being able to quantify a possible indirect effect of indoor air on food via sill dust.

Dunson's concern that we may have relied too heavily on Clayton et al. (2002)'s pathways structure is certainly valid. While the NHEXAS data's weaknesses (i.e., large fractions of missing and censored observations) necessitated the use of external information in modeling building, there are possibly more flexible model specifications. Dunson's suggestion that we allow uncertainty in the ADG is intriguing. This proposal amounts to putting a prior on the ADG and treating the presence or absence of links as a part of the randomness. This is an interesting area of research that is still in its formative stages (e.g., Ellis and Wong 2008; Hunter et al. 2008). It represents an elegant way to incorporate uncertainty on the presence or absence of links, but how to weight the ADGs to represent scientific meaningful pathways is an open problem as far as we know. Further, criteria would have to be imposed if we wanted to infer an "optimal" ADG from its posterior distribution. With all that is happening in our Bayesian hierarchical model, we prefer our "nuts and bolts" approach to adding or removing model components.

4.2 Model checking

The use of Figures 8 and 11 are purely exploratory. While only one posterior realization is represented in each, a decrease in the variation in the standardized residuals is apparent in Figure 11. This dilemma of how to demonstrate pattern in high-dimensional posterior distributions also arises in a similar way in statistical image analysis. One way around this problem is to establish various operating characteristics that summarize the behavior of an image. Likewise, here we would like to define some informative summaries of the six plots shown in each of Figures 8 and 11. The paper by Gelman et al. (2005), which we should have referenced when we discussed posterior predictive checks, requires the specification of a random mechanism for how the data became missing or censored. However, we have incomplete information about this mechanism from NHEXAS's documentation and relevant papers (Robertson et al. 1999; O'Rourke et al. 1999). While the missing-at-random assumption may be questionable, we do not have enough detail on the mechanism to incorporate it into our model.

4.3 Model comparisons

We can expand a little on our strategy given at the beginning of Section 4.3 of our article. First, all the models we build are ADGs, and when we add a model component it is always done using directed edges that preserve the ADG structure. Each ADG is made up of vertices that represent the (hidden) process model and vertices that represent the data (observed, censored, missing). If there are two process vertices, P and Q, with a purported directed edge from P to Q, our goal is to use the data Y^P and Y^Q, respectively, to see if the link is really there. If there is no link, then $[Q|Y^Q] = [Q|Y^Q, Y^P]$. Hence, if the posterior distribution of Q (or of any component of Q) changes depending on whether we condition on Y^Q versus when we condition on (Y^Q, Y^P), then we have evidence of learning and the link should be kept. We did this many times for many choices of P (e.g., global soil) and Q (e.g., LEB) and on occasions we also looked for evidence of learning by comparing $[P|Y^P]$ to $[P|Y^P, Y^Q]$. We compared posterior summaries to judge whether learning had occurred, but beyond this we have no formal way of scoring the various models.

4.4 Limitations of policy-related conclusions

We wholeheartedly agree with Barr and Dominici on the importance of clear articulation of the scientific and policy questions underlying any applied statistical analyses. In our study, our goals were more scientific in nature than policy oriented, although we acknowledge that the two can rarely be separated completely. The primary scientific motivation for our study was to explore the extent to which supplemental data on the spatial variation in background levels of arsenic in environmental media could be used to refine our understanding of arsenic-exposure pathways and of the geographic variation in exposure to arsenic. We readily acknowledge that our discussion of modeling issues dominated the discussion of our scientific conclusions, which appear only briefly

in Section 5. In part, this was the result of our ability to make only weak scientific conclusions – with regard to the inclusion of global environmental media, we can conclude that it only helps discern geographic variation in exposure to arsenic slightly and does not provide much additional understanding of arsenic-exposure pathways beyond what can be derived from NHEXAS data alone.

4.5 Units

Our not giving the units in Figure 9 was an oversight. The measurement units for all the quantities in the LEB model are the natural log of $\mu g/l$, as discussed in Section 2.2 of our article. The process variances in Figure 10 are measured in the square of these units.

5 Reproducible research

We agree with Barr and Dominici, and Peng et al. (2006), that reproducible research is a noble goal. Time permitting, we hope to make our code available. As we discuss at the end of the article, it helps to produce computer code in bite-size pieces that can be checked separately. This also encourages code re-use, which in turn benefits the research community. Additionally, making code available to reconstruct derived datasets from original data sources is an important component of reproducible research.

References

ATSDR (2005). *Toxicological Profile for Arsenic*. Draft Report of the Agency for Toxic Substances and Disease Registry. Public Health Services, U.S. Department of Health and Human Services, Washington, DC.

Brooks, S. P. and Gelman, A. (1998). "General methods for monitoring convergence of iterative simulations." *Journal of Computational and Graphical Statistics*, 7: 434–455.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. New York, NY: Chapman and Hall/CRC.

Clayton, C. A., Pellizzari, E. D., and Quackenboss, J. J. (2002). "National Human Exposure Assessment Survey: Analysis of exposure pathways and routes for arsenic and lead in EPA Region 5." *Journal of Exposure Analysis and Environmental Epidemiology*, 12: 29–43.

Cressie, N., Buxton, B. E., Calder, C. A., Craigmile, P. F., Dong, C., McMillan, N. J., Morara, M., Santner, T. J., Wang, K., Young, G., and Zhang, J. (2007). "From sources to biomarkers: A hierarchical Bayesian approach for human exposure modeling." *Journal of Statistical Planning and Inference*, 137: 3361–3379.

Ellis, B. and Wong, W. (2008). "Learning causal Bayesian network structures from experimental data." *Journal of the American Statistical Association*, 103: 778–789.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis, 2nd edition*. Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A. and Rubin, D. B. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science*, 7: 457–472.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D. F., and Meulders, M. (2005). "Multiple imputation for model checking: Completed-data plots with missing and latent data." *Biometrics*, 61: 74–85.

Hunter, D., Goodreau, S., and Handcock, M. (2008). "Goodness of fit for social network models." *Journal of the American Statiatical Association*, 103: 248–258.

O'Rourke, M. K., Van de Water, P. K., Jin, S., Rogan, S. P., Weiss, A. D., Gordon, S., Moschandreas, D., and Lebowitz, M. (1999). "Evaluations of primary metals from NHEXAS Arizona: Distributions and preliminary exposures." *Journal of Exposure Analysis and Environmental Epidemiology*, 9: 435–445.

Peng, R., Dominici, F., and Zeger, S. (2006). "Reproducible epidemiological research." *American Journal of Epidemiology*, 163: 783–789.

Robertson, G., Lebowitz, M., O'Rourke, M., Gordon, S., and Moschandreas, D. (1999). "The National Human Exposure Assessment Survey (NHEXAS) study in Arizona – introduction and preliminary results." *Journal of Exposure Analysis and Environmental Epidemiology*, 9: 427–434.

Whitmore, R., Byron, M., Clayton, C., Thomas, K., Zelon, H., Pellizzari, E., Lioy, P., and Quackenboss, J. (1999). "Sampling design, response rates, and analysis weights for the National Human Exposure Assessment Survey (NHEXAS) in EPA Region 5." *Journal of Exposure Analysis and Environmental Epidemiology*, 9: 369–380.