Bayesian Analysis (2009) **4**, Number 1, pp. 37–40

# Comment on Article by Craigmile et al.

Christopher David Barr* and Francesca Dominici†

**Abstract.** We congratulate the authors on their development of Bayesian hierarchical models that simultaneously overcome many of the challenges inherent to analyzing exposure pathways to Arsenic. Drs. Craigmile, Calder, Li, Paul and Cressie (herein referenced by CCLPC) address a range of obstacles with a sound mix of sophisticated models, useful plots and common sense. Their paper is well written throughout, making important illustrative, methodological and applied contributions. We thank the editor for inviting us to discuss this stimulating paper, and will begin by calling special attention to some particular highlights of the analysis by CCLPC. After which, we will mention a few aspects we believe can be further improved.

## 1 Bayesian Hierarchical modeling for complex environmental data

The popularity of Bayesian hierarchical models is growing quickly (Banerjee, Carlin and Gelfand (2004) provide a nice introduction from a spatial-temporal perspective), and the paper of CCLPC is a shining example of the breadth and flexibility of this important class of models. Bayesian Hierarchical models have been used to address many of the issues faced in this paper, including spatial misalignment, censored data, missing data, spatial clustering, and direct and indirect effects. The complexity of dealing with even a couple of these issues simultaneously can be daunting, and we applaud CCLPC for their trenchant modeling of the complex Arsenic exposure data.

We appreciate the modular approach taken by CCLPC both in modeling and presentation. Especially nice is the liberal use of graphs and sparing use of tables (see Gelman et al. (2002)) to summarize data exploration (e.g. Figure 4), model construction (e.g. Figure 3), model fitting (e.g. Figure 8), and broader findings (e.g. Figure 9). The authors comparison of the LEB and LEB + GLE models is a good example of the benefits of building and diagnosing models piece-by-piece. We only wish CCLPC had more room to discuss the global models as a group before evaluating their contribution to the larger framework.

*Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, mailto:cbarr@jhsph.edu

†Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, mailto:fdominic@jhsph.edu

 DOI:10.1214/09-BA401A

## 2 Balancing simplicity and sophistication

Over-fitting and limited interpretability are common threats in Bayesian modeling, and become increasingly worrisome as the model grows in complexity. We congratulate CCLPC on the numerous steps they have taken towards addressing these issues, and continue with a few suggestions for making complex analyses such as theirs even more focused.

An important benchmark for the success of an applied paper is its usefulness in a policy setting. While this paper is meant as an illustration for a statistical audience, we believe it would have benefited from a clearer articulation of the scientific and policy questions being addressed. In Section 2.1 the authors emphasize the scientific importance of identifying "geographical variation in exposure to hazardous chemicals". Figure 13 addresses this issue somewhat, but the topic receives only limited discussion in their scientific summary (Section 5). Another important concern is interpretability of coefficients. The authors discussion of direct and indirect effects (Section 2.1, paragraph 2) is helpful, as is Figure 9, which reports results for slope parameters for direct pathways to Urine. However, we wish space permitted the inclusion of units within Figure 9 or its caption.

This paper also raises many good discussion points concerning defensible construction and estimation of Bayesian hierarchical models. When faced with challenging data and complex relationships, as in the paper of CCLPC, "grounding" the modeling process can be especially difficult.

In their paper, CCLPC take modeling cues from the existing literature (e.g. Section 3.1) and discuss the appropriateness of such suggestions within the context of their analysis. This is an excellent practice, and we believe the use of simple measures (e.g. Figure 4) is sensible for making the link between existing and new models. In fact, we would have liked CCLPC to continue the discussion of Section 3.1 with more comments on how existing models did not seem as well suited (e.g. zero correlation between food and indoor air).

Interestingly, another strategy CCLPC use to "ground" their model is the model itself. By growing their model and code (in what appears to be one direction), they stake the validity of their conclusions on a series of modeling components and assumptions, which are dependent on those earlier in the modeling process to an alarming extent. In Section 4.2, for example, CCLPC "initially fixed most of the model parameters at reasonable values", but we worry that the extent to which initial values are "reasonable" may be subject to the order in which the components are assembled. Considering different orders of assembly is one response to this challenge. Another option is to use a simpler model, such as multivariate regression to validate the more complex (and potentially more accurate) hierarchical Bayesian model at various stages.

# 3 Reproducible research

In environmetric analyses it can be difficult to determine which variables to include in a model, and by which functional forms to connect them. During exploratory investigations the data should play a central role in answering these questions. However, during inferential studies, it is important to the scientific process that models be formulated to the greatest extent possible before seeing the data. We congratulate CCLPC on the steps they took to use existing literature to motivate their models, and we encourage others to pursue a priori model formulation when possible.

Just as we support the reproduction of existing research, we champion good habits in making ones own research reproducible (Peng et al. (2006)). The authors have facilitated this through efforts to document their data management practices, and we hope they will also make code available when time permits.

# References

Gelman, A., Pasarica, C., and Dodhia, R. (2002). "Lets practice what we preach: Turning tables into graphs." *American Statistician*, 56: 121–130.

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