Discussion of the paper by Pardoe and Weidner

It gives me great pleasure to discuss the interesting application of multilevel random effect modelling to crime data given in Pardoe and Weidner. Random effect modelling is increasing in popularity in virtually all application areas as applied researchers realise the importance of accounting for the structure of the data they have collected and any non-independence that exists. Even within the area of crime and justice, multilevel models are used for many problems as diverse as assessing determinants of inmate violence (Huebner, 2003) to looking at disorder in city neighbourhoods (Sampson and Raudenbush, 1999). Pardoe and Weidner give an extensive list of further uses within this application area. Much of this increase in interest can be attributed to user-friendly software packages that are accessible to applied researchers. From a classical point of view packages like HLM (Raudenbush et al., 2001) and MLwiN (Rasbash et al. 2000) have offered users methods for finding maximum likelihood estimates for random effect models for several years, and with the incorporation of some random effect modelling in virtually all general statistics software packages we can expect far more analyses being published that use such methods.

The incredible increase in the speed of modern computers has allowed Monte Carlo Markov chain (MCMC) methods to not only become computationally feasible but to produce estimates in a reasonable time for realistic applied problems. The advent of the WinBUGS package (Spiegelhalter et al. 2003) and the incorporation of some MCMC methods in packages like MLwiN (Rasbash et al. 2000) has introduced MCMC methods, and to a certain extent Bayesian statistics to a large community of applied users. MCMC methods have advantages over most algorithms that give likelihood-based estimates, in particular in the way that algorithms can be easily 'bolted together' to fit additional model features. Pardoe and Weidner illustrate this with their incorporation of missing data imputation within their model structure. The other main advantage of MCMC methods is that as they are simulation-based they output chains of values from the posterior distribution of their unknown parameters. This allows easy derivation of posterior distributions of functions of parameters and this property is illustrated well by Pardoe and Weidner in their figures of odds ratios.

In my discussion I would like to talk about four statistical topics touched on in this paper; firstly the authors talk briefly about sensitivity analysis and so I have a few comments on the sensitivity of prior distributions; second the authors introduce an interesting univariate missing data step into their model and I will describe some alternatives; thirdly the authors present a single model with some justification of exclusion of certain terms but no model comparison which merits some comment. Finally I will discuss the authors use of interaction terms in the model.

1 Prior sensitivity

One major issue when using the Bayesian approach is the incorporation of prior distributions for the unknown parameters in the model. The Bayesian approach has the advantage that if we have additional information about parameter values that we can incorporate this information into our modelling via prior distributions. Some Bayesians will argue that we always have prior knowledge about the unknown parameters in our model however often this information may be in fact that we think all (or at least a large range of) values are equally likely for the parameter. Thus there is a need for 'diffuse' or flat priors to express our ignorance about the parameters prior to collecting the data. For certain types of parameters, for example variance matrices, there is still some debate on what is a sensible prior to use to represent ignorance. Pardoe and Weidner's model contains 13 sets of correlated random effects, α_j and these result in a 13*13 variance matrix, Γ^{-1} which they give an inverse Wishart prior.

$$\Gamma \tilde{} \ \text{Wishart}(R, K)$$

When discussing prior distributions they state that 'with small samples this choice can be critical, but with larger samples (such as in this application) the choice is less crucial'. It should of course be noted here that the important sample size here is the number of counties (39) and so I would contend that in this example this is small given that a 13*13 matrix is being estimated from just these 39 counties and hence the choice of prior is important. Browne and Draper (2000) look at simulations of a continuous multilevel model which contains a 2*2 variance matrix and compare two inverse Wishart priors showing some of the biases that can occur in small samples. Pardoe and Weidner give values of 10 along the diagonal of R and zeroes elsewhere which should be justified. I would be interested in seeing how much is in fact gained by including all 13 sets of random effects and whether a simpler model here would be more appropriate given the fairly small number of counties. Other approaches to finding a prior for variance matrices include the approach of Barnard, McCulloch and Meng (2000) that considers instead the decomposition of the prior parameters into a set of standard deviations and correlations which the researcher may find easier to attribute prior estimates to. It appears in fact that estimation of the variance matrix Γ^{-1} is not of great interest to the authors since they do not quote estimates for the matrix anywhere in the paper. Such estimates can be used to find out the importance of the county in accounting for variability of the response through the variance partition coefficient (see Goldstein et al. 2002). I did however like the use of figures (3-7) to explain the main effects and cross level interactions.

2 Missing data imputation methods

One advantage of the authors MCMC approach is that they can build in an imputation step directly into their algorithm to account for missing data. This means that rather than throwing away any records containing some missing data or running some imputation to create one or many 'complete' datasets prior to estimating their model, the authors can instead treat the missing data as additional parameters to be estimated by the model. In their dataset the authors have 5 binary predictor variables that contain missing data which range in quantity from 0.1% missing for the predictor IMALE to 32.3% missing for the predictor IPPRIS. For each predictor they fit a logistic regression model based on a choice of predictors (from the remaining 7 complete predictors) that appear to be correlated with the particular predictor. Although it is sensible not to have predictors with missing data being used in the prediction of other predictors there still may exist correlations between these predictors and the authors avoid attempting to model this in their modelling. Looking at the predictors (containing missing values) the largest correlation is 0.27 between IACTCJS and IDETAIN which might improve prediction of missing values. To include correlations between the missing predictors the authors could use the latent variable approach of Chib and Greenburg (1998) to fit a multivariate probit model for the missing responses. Alternatively they could treat the 5 binary predictors as a multinomial response with 32 (2^5) possible responses and then the missing responses would each take values from a subset of the possible responses depending on which predictors are missing. This approach would I suspect however be more difficult to implement. One reason the authors may not have considered fitting a multivariate probit model is that this is difficult to fit in the WinBUGS software they are using.

A final point on the missing data that needs considering is the differences that exist between the counties in the dataset. Examining the data it is clear that the predictor variables have different distributions in each county, for example the percentage of IBLACK=1 ranges from 15.3% to 90.0% at the county level. However the amount of missing data is also dependent on the county in question, for example in county 12 the predictor IPPRIS is totally missing and so is being predicted from the other counties where it is present. Given that some of the counties also have 100% of their non-missing observations having IPPRIS=1 but only a small number of observations non-missing it would be interesting to find out why in these cases the variable is missing. It would also be worth investigating if there are particular county policies on data collection for certain predictor variables that would influence whether an observation is missing.

3 Model Selection

I would like to have seen more information given on why the authors came up with their final model. These days there are many MCMC methods that include model selection in the modelling process. For example reversible jump MCMC (Green 1995) can be used to perform model averaging over a family of plausible models. There also exist diagnostics for example the DIC diagnostic of Spiegelhalter et al. (2002) that allow comparisons to be made between alternative models. Although I commend the authors for their model assessment (section 5 of the paper) which is often neglected I would have preferred to have also seen some more evidence of model selection and some evidence that the large number of random effects in their model is really required.

4 Interaction terms

The authors model contains a phenomenal number of interaction terms. In fact only five possible 'cross-level' interactions (interactions between a predictor variable at the individual level and a predictor variable at the county level) are discounted and the reasons for these omissions are given in the paper. I was surprised by the number of interactions included, particularly as very few of them appear to be statistically significant. On this point I question why the authors highlight in bold posterior means that are bigger than their standard deviations; why not choose the more common two standard deviations? My other main surprise in the authors approach to interactions is the neglect of standard interactions. For example have the authors tested for a simple interaction between predictor variables such as IBLACK and IPPRIS? This would seem equally as important to me as the cross-level interactions that the authors focus on.

5 References

Barnard, J., McCulloch, R. and Meng, X-L (2000). Modelling Covariance Matrices in Terms of Standard Deviations and Correlations, with Application To Shrinkage. *Statistica Sinica*, 10, 4.

Browne, W.J., and Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics* **15**: 391-420.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85** (2): 347-361

Goldstein, H., Browne, W.J. and Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics* **1**: 223-232.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** (4): 711-732

Huebner, B.M. (2003). Administrative determinants of inmate violence: A multilevel analysis. *Journal of Criminal Justice* **31**: 107-117.

Rasbash, J., Browne, W.J., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I. and Lewis, T. (2000). *A User's Guide to MLwiN* (2nd edition). London: Institute of Education.

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F. and Congdon, R. (2001). *HLM 5: Hierarchical Linear and Nonlinear Modelling* (2nd edition). Chicago: Scientific Software International.

Sampson, R. J. and Raudenbush, S.W. (1999). Systematic social observation of public spaces: A new look at disorder in urban neighbourhoods *American Journal of Sociology* **105**: 603-651.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistics Society, Series B* **64**: 583-640.

Spiegelhalter, D.J., Thomas, A., Best, N.G., and Lunn, D. (2003) *WinBUGS Version 1.4 User Manual*. Cambridge, UK: MRC Biostatistics Unit.