Bayesian Analysis (2008) **3**, Number 2, pp. 263–268

Rejoinder

P. G. Blackwell* and C. E. Buck†

We would like to thank all of the discussants for their comments which place our work in context and help us to step back a little. This is an extremely timely moment to do this, since we are about to embark on a NERC funded project that aims to develop the models and methods to provide the next internationally-agreed estimates of the calibration curves which are due to be released in 2010. We respond to the written comments from the discussants under three broad headings as follows.

1 Improved modelling of the physical processes

One of the themes in the comments offered by the discussants is that we should seek statistical models that have greater foundation in the processes observed in the physical world. As noted in our discussion section, inclusion of further prior information about $\mu(\cdot)$ is something that we certainly intend to explore in the production of the next curve.

Millard cautions against circularity in the use of the data when modelling periodicity in the curve. Certainly we had no intention of building in cycles of particular lengths; however, allowing for the possibility of periodicity, and seeing what emerges, would seem a natural (and valid) way forward. Alongside this, we will also explore the use of heavier tailed distributions for our random walk prior in order to be sure that we are not over-smoothing; this also helps ensure that we are not over-reacting to outliers, an issue revisited below. For this purpose, the suggestion by Haslett and Parnell of an infinitely-divisible model such as the Normal Inverse Gaussian is very helpful. Applying this in practice will require some care, since we require a method that picks up features supported by multiple data points and that does not too readily create 'wiggles' based on single data points.

Closely allied to these suggestions about modelling the underlying processes is a suggestion from Millard to use more realistic models for the uncertainties on the data, in particular those on the calendar age scale. Exploration of such complexities was not an option given the implementation restrictions for IntCal04, but Millard is quite right that our assumption of normal errors everywhere is simplistic and needs further investigation. When we come to do this, Millard's own work to develop Bayesian models for uranium series dating (Millard 2004, 2006) will be invaluable.

Arguably, the least satisfactory of all of our modelling for IntCal04 was that relating to data derived from sedimentary sequences whose ordering is known and must be incorporated in the analysis. The move to an MCMC implementation allows a great deal more choice about how we handle this aspect of the problem, and we plan to

*Department of Probability and Statistics, University of Sheffield, UK http://paul-blackwell.staff.shef.ac.uk/

†Department of Probability and Statistics, University of Sheffield, UK, http://caitlin-buck.staff.shef.ac.uk/

 DOI:10.1214/08-BA309REJ

investigate a range of options. Haslett and Parnell draw attention to their recent work on modelling sequences of this type which clearly has its attractions. It becomes even more attractive when we note that some of the potential data providers (particularly those with long marine records) have to use some kind of deposition model in order to provide calendar date estimates for all depths in their sequence, since only some depths provide material that can be dated directly. At present, the most common approach is to provide calendar date estimates via interpolation which typically implies linear deposition rates throughout long periods of time. Haslett and Parnell's method offers many advantages over these existing ad hoc approaches and is fully Bayesian so could be incorporated into our framework for curve estimation. This would, however, add considerable complexity to the structures currently in use and increase computation times. Again, we will investigate the range of options, but would not readily move to more complex models unless they clearly provided considerable benefit for the inferences we can make and were computationally feasible.

Something else that clearly seems unsatisfactory to some readers, since Millard and one of our referees both raised it, is the apparent smoothness of the calibration curve we produce. Although we have already added a little to the final paper in response to the referee's comment, it seems worth elaborating here. There is a crucial distinction to be made between the smoothness of our posterior mean for $\mu(\theta)$, viewed as a function of θ, and our posterior beliefs about the smoothness of $\mu(\cdot)$ itself, which are not shown explicitly in the paper but are certainly implicit in, for example, our use of the curve to produce Figure 4. In short, the smoothness of the estimate is not an estimate of the smoothness. While there remains plenty of scope for refining our prior modelling of the smoothness of the curve, perhaps along the lines suggested by Haslett and Parnell, the absence of wiggles that were in earlier versions of the curve may simply mean that they are not justified by the data.

2 Improvements in data modelling and data quality issues

Haslett and Parnell are correct that our mention of covariance in Section 6.6 is too narrowly focused; it would have been better simply to refer to dependence throughout, as we did in our Discussion. Our point is, of course, that *all* the information on the joint distribution of *all* points on the curve can be used in the calibration of single or multiple samples. That is exactly what was done in producing the first posterior distribution in Figure 4. Millard points out that allowing for dependence does not always make a large difference to the calibration of single samples, but we should expect that the impact will depend, as he speculates, on the density of the data—actually, on the density of the data relative to the uncertainty in the uncalibrated observation—as well as on the steepness of the curve, the dependence between observations, and many other details. His fourth example and our Figure 4 suffice to show that allowance *should* be made for dependence, unless there is a clear argument that an approximation based on assuming independence will be adequate. In general the effect is hard to predict, although increased uncertainty about individual dates and wiggle-matching, and decreased uncertainty about differences or ranges, seem likely.

In the final paragraph of their Section 2, Haslett and Parnell raise issues of data quality and ask for our thoughts on the detection and handling of outliers. As the discussants hint, there are several types of unquantified error (outlier) that we need to be aware of in radiocarbon dating:

1. relatively small and ubiquitous 'outliers' that arise from errors in the internal radiocarbon laboratory procedures, but that are not accounted for in the the laboratory's estimates of its internal error;

2. large and fairly common outliers (Scott, 2003, estimates roughly 1 in 20) that arise for a range of reasons, including contamination of samples either in the field or inside the laboratory;

3. extremely large outliers, that occur due to contamination and/or misattribution of samples in the field (e.g. via the dating of collections of unidentified organic material from lake sediments);

4. outliers caused when whole sequences of related dates (e.g. from lake or ocean sediments) have, together, been misunderstood or misinterpreted, for example due to inaccurate assessment of the amount of local upwelling of old carbon or to an error in wiggle matching.

Statisticians have advocated a range of approaches to identifying and handling each of these problems. Marian Scott has worked with the radiocarbon laboratories for many years, helping them systematically to assess the quality of their laboratory practices and to quantify any extra uncertainty not accounted for in the standard deviation (s_j for determination j, in our notation) that they supply to their customers. Her work has shown that very few laboratories have problems with systematic offsets in their radiocarbon determinations but that many laboratories do under-estimate their uncertainties.

One output from Scott (2003) is something that she calls a laboratory error multiplier. This is an indication of the amount by which individual laboratories need to inflate their error estimates if they are to capture the extra sources of uncertainty not accounted for in s_j. In our paper we denote the error multiplier for laboratory l by k_l, and thus define the total variance on a radiocarbon determination as $\sigma_j^2 = s_j^2 k_{l_j}^2$ if determination j comes from laboratory l_j. Scott found that k_l is close to 1 for quite a number of laboratories, but that the median value is around 2, "suggesting that the quoted uncertainties are, in general, too small" (Scott, 2003, p. 326). Although Scott's work is published in such a way that individual laboratories cannot be identified by name, all of the laboratories are provided with an estimate of their own laboratory error multiplier and the laboratories that supply data to the IntCal database have been happy to report their k_l value along with their other data.

For detecting and handling both types of more substantial outliers that arise from contamination or misattribution of samples, there are natural Bayesian techniques that are well established—there are too many references to list, but see Box and Tiao (1968)

and citations thereof—but their uptake in the calibration literature has been very limited. Christen (1994) provided a fully Bayesian solution to the problem of moderate sized errors which occur due to relatively minor contamination or to mobility of organic material in the ground. Christen's approach defines a radiocarbon determination as an outlier if it needs a shift in its radiocarbon determination in order to make it consistent with the rest of the samples being interpreted alongside it (typically those arising from a well-stratified sequence or a single archaeological phase). Christen's method is now widely used and implemented in packages such as OxCal and BCal. In an extension to Christen's method, Haslett and Parnell (2008) suggest that some radiocarbon determinations are so outlying that they cannot be brought into line with neighbouring ones and should simply be left out of the analysis altogether. Rather than rejecting such determinations in an ad hoc fashion as has been done in the past, Haslett and Parnell (2008) recommend a fully Bayesian extension to the method provided by Christen. They label determinations that simply need a shift to bring them in-line with others as Type 1 outliers and define Type 2 outliers as determinations for which none of the calendar age probability distribution satisfies the conditions of the model in use (in their case study monotonicity between depth in a sequence and calendar age). Although not yet widely adopted, Haslett and Parnell's extension is coded into their new R package for the construction of age-depth models (known as Bchron, available via CRAN) and seems likely to become popular with the user communities for which it was designed (i.e. those seeking to construct chronologies for past environmental sequences from peat and lake sediments).

A greater difficulty is caused when whole sequences of related dates have, together, been misunderstood or misinterpreted, as arose in another part of our calibration work, relating to the time period before about 26,000 years BP. We chose not to discuss this aspect of our work at the Case Study meeting because it did not lead to an internationally-agreed curve estimate. Nonetheless, it was an important part of the project. For that time period, the IntCal group had several long sequences of data; the problem was that these sequences did not all meet the quality control criteria for inclusion in the IntCal database and had obvious offsets one to another. Our initial response to this was to talk with the data providers and to seek obvious explanations for the offsets we observed. However, no such explanations could be found and so we devised a random effects extension to the models used for estimating IntCal04, which allowed us to take account of systematic offsets in sequences of data at the same time as estimating the underlying curve (Buck and Blackwell 2004). By applying this method, we were able to demonstrate that some of the sequences of data in the NotCal database required offsets as large as 2000 years to bring them in-line with the other data and, hence, with our estimate of the underlying curve (van der Plicht et al. 2004). Although by doing this, we were able to provide an estimate of the curve and a predictive distribution for future observations, when the other members of the IWG saw the scale of the random effects required they felt that we should not attempt to release this as an internationally-agreed curve and so the label NotCal was applied.

More generally, while complex models do present challenges for the wider issues of the checking and criticism of both data and models, there are some techniques avail-

able, taking advantage of the coherence of the Bayesian approach and the flexibility of modern computational tools. An example of this is illustrated in Figures 2 and 3 in the main paper. These figures show how prior and posterior distributions match up for the calibration data, and give a natural starting point for thinking about both the influence of individual observations and the possibility of gross errors. Of course, this approach exploits the fact that uncertainty on individual observations is already considered in this analysis. Another powerful technique that is more widely applicable is to compare individual observations with their posterior predictive distributions (e.g. Gelman et al. 2004), as implemented for a rather more complex model in Blackwell (2003). Finally, a simple technique that nevertheless has great potential for model criticism and interpretation is to display a small sample of realisations from the prior and/or posterior for a model, or an appropriate part of a model; an example of this, showing the effects of different model structures, is given by Blackwell and Møller (2003).

3 Suggestions for implementational improvements

Haslett and Parnell suggest an alternative implementation to the one we describe, alternating block updates of $\mathcal{M} = \mu(\Theta)$ with block updates of Θ. It is certainly true that the full conditional for $\mu(\Theta)$ has a simple form, and block updating in that case may be profitable. However, the full conditional for Θ does not have such a simple form, because of the dependence on θ_j of both the mean and variance on the right hand side of equations (5) and (6) in our paper. An actual Gibbs update here does not seem possible; a Metropolis-within-Gibbs algorithm that updates Θ as a block, separately from $\mu(\Theta)$, would be possible, but it does not seem clear that it would be more efficient than the current approach. One advantage of the current approach is that μ_j and θ_j, which are strongly dependent, are updated together; separating them may impede mixing. Of course, adding the suggested updates to our approach may improve mixing.

References

Blackwell, P. G. (2003). "Bayesian inference for Markov processes with diffusion and discrete components." *Biometrika*, 90: 613–62. 267

Blackwell, P. G. and Møller, J. (2003). "Bayesian inference for deformed tessellation models." *Advances in Applied Probability*, 35: 4–26. 267

Box, G. E. P. and Tiao, G. C. (1968). "A Bayesian approach to some outlier problems." *Biometrika*, 55: 119–129. 265

Buck, C. E. and Blackwell, P. G. (2004). "Formal Statistical Models for Estimating Radiocarbon Calibration Curves." *Radiocarbon*, 46(3): 1093–1102. 266

Christen, J. A. (1994). "Summarizing a set of radiocarbon determinations: a robust approach." *Applied Statistics*, 43(3): 489–503. 266

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC Press. 267

Haslett, J. and Parnell, A. (2008). "A simple monotone process with application to radiocarbon dated depth chronologies." *Journal of the Royal Statistical Society - Series C*. To appear. 266

Millard, A. R. (2004). "Taking Bayes beyond radiocarbon: Bayesian approaches to some other chronometric methods." In Buck, C. E. and Millard, A. R. (eds.), *Tools for Constructing Chronologies: crossing disciplinary boundaries*, 231–248. London: Springer-Verlag. 263

— (2006). "Bayesian analysis of Pleistocene Chronometric Methods." *Archaeometry*, 48(2): 359–375. 263

Scott, E. M. (2003). "The fourth international radiocarbon inter-comparison." *Radiocarbon*, 45(2): 135–408. 265

van der Plicht, J., Beck, J. W., Bard, E., Baillie, M. G. L., Blackwell, P. G., Buck, C. E., Friedrich, M., Guilderson, T. P., Hughen, K. A., Kromer, B., McCormac, F. G., Bronk Ramsey, C., Reimer, P. J., Reimer, R. W., Remmele, S., Richards, D. A., Southon, J. R., Stuiver, M., and Weyhenmeyer, C. E. (2004). "NotCal04—comparison/calibration ^{14}C records 26–50 cal kyr BP." *Radiocarbon*, 46(3): 1225–1238. 266