Bayesian Analysis (2008) **3**, Number 2, pp. 225–248

Estimating radiocarbon calibration curves

P. G. Blackwell* and C. E. Buck†

Abstract. In addition to being crucial to the establishment of archaeological chronologies, radiocarbon dating is vital to the establishment of time lines for many Holocene and late Pleistocene palaeoclimatic studies and palaeoenvironmental reconstructions. The calibration curves necessary to map radiocarbon to calendar ages were originally estimated using only measurements on known age tree-rings. More recently, however, the types of records available for calibration have diversified and a large group of scientists (known as the IntCal Working Group—IWG) with a wide range of backgrounds has come together to create internationally-agreed estimates of the calibration curves. In 2002, Caitlin Buck was recruited to the IWG and asked to offer advice on statistical methods for curve construction. In collaboration with Paul Blackwell, she devised a tailor-made Bayesian curve estimation method which was adopted by the IWG for making all of the 2004 internationally-agreed radiocarbon calibration curve estimates. This paper reports on that work and on the on-going work that will eventually provide models, methods and software for rolling updates to the curve estimates.

1 Introduction

1.1 Radiocarbon dating

Carbon is one of the most abundant elements and, due to its capacity to bond with other elements, it is an essential part of all living organisms. As a result, in the late 1940s, when Willard Libby and colleagues discovered a method for estimating the age of carbon-containing materials there was great excitement in a number of disciplines including archaeology, geology and palaeontology.

Libby's method relies on the fact that one of the isotopes of carbon—^{14}C—is radioactive. While they are alive, all plants and animals exchange carbon (including ^{14}C) with the atmosphere. This means that, during life, living organisms have levels of ^{14}C that are in approximate equilibrium with the atmosphere. Upon death, however, plants and animals cease metabolising and so no new carbon is taken up. At this point, the amount of ^{14}C in any surviving tissue begins to decrease exponentially, via radioactive decay.

If we assume, as Libby and colleagues did, that the ratio of ^{14}C to other carbon isotopes has been constant over time then all we need to date a sample of (long) dead carbon are: an estimate of the proportion of ^{14}C in modern carbon, an estimate of the half-life of ^{14}C and a method to estimate the current proportion of ^{14}C in the sample

*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-BA309

itself. Libby and colleagues provided all of these and so radiocarbon dating appeared both simple and reliable. It also appeared particularly useful to archaeologists since Libby's estimate of the half-life was 5568±30 years; i.e. about right to help date most of the human past. The current estimate of the half-life is 5730±40 years, but to avoid confusion (given the need for calibration explained below) all radiocarbon laboratories still use the Libby half-life to calculate radiocarbon ages.

1.2 The need for calibration

As discussed in some detail in Aitken (1990) and Bowman (1990), quite quickly after radiocarbon dating was discovered, it became clear that Libby's assumption of constant ^{14}C levels in the atmosphere does not hold. The level is affected by very many complex factors that have proven impossible to model mechanistically. These include natural factors such as: solar cycles, solar storms, geomagnetic variations in the earth and unpredictable up-welling of old carbon from substantial reservoirs such as oceans. The level has also been impacted by human activity; for example, it increased substantially due to atomic bomb testing in the 1950s and has dropped again more recently due to release of old carbon in fossil fuels.

As a consequence, radiocarbon dating is only viable if we can obtain an estimate of the varying level of ^{14}C back through time and can thus plot the function that links radiocarbon ages to calendar ages. Put loosely, we need a calibration curve. Since the factors affecting variations in ^{14}C level are too complex to model mechanistically the only way to get such a curve is to estimate it from data.

1.3 The calibration curve

Given a radiocarbon determination X_0 for a sample, we wish to assess the true age θ_0 of the sample by taking into account the information from other samples of known (or approximately known) age, our 'calibration data', which are paired observations of radiocarbon determinations and calendar ages $\mathcal{X} = (X_1, \ldots, X_N)$ and $\mathcal{T} = (T_1, \ldots, T_N)$; we denote the true ages of the calibration samples by $\Theta = (\theta_1, \ldots, \theta_N)$. Thus we are interested in

$$p(\theta_0|\mathcal{X}, \mathcal{T}, X_0).$$

An object of age θ has 'true' radiocarbon level $\mu(\theta)$ say, so that $\mathrm{E}(X_0|\theta_0) = \mu(\theta_0)$, and we write $\mathcal{M} = (\mu_1, \ldots, \mu_N) = (\mu(\theta_1), \ldots, \mu(\theta_N))$ for the corresponding 'true' levels for the calibration data. Furthermore X_0 depends on $\mathcal{X}$ and $\mathcal{T}$ only through $\mu(\cdot)$. Thus, although it was not articulated in this way originally, at the core of the radiocarbon dating method is a process that can be summarised by the function $\mu(\cdot)$, the so-called 'calibration curve'. We can write

$$\begin{aligned} p(\theta_0|\mathcal{X}, \mathcal{T}, X_0) &= \int p(\theta_0|\mu, X_0) p(\mu|\mathcal{X}, \mathcal{T}, X_0) \mathrm{d}\mu \\ &= \int p(\theta_0|\mu, X_0) p(\mu|\mathcal{X}, \mathcal{T}) \mathrm{d}\mu, \end{aligned} \tag{1}$$

making the realistic assumption that X_0 on its own tells us essentially nothing about the calibration curve. This paper primarily concerns inference about $\mu(\cdot)$. Calculation of $p(\theta_0|\mu, X_0)$ and the integral in (1), and hence the actual calibrating of an individual date, are addressed only briefly in Sections 2, 6.6 and 7.

2 Historical background

The work reported in this paper builds on research by a large number of people (some very famous and many unrecognised) in several disciplines who have contributed to making radiocarbon dating ubiquitous. The first of these was Willard Libby (1908 to 1980) whose role in developing radiocarbon dating (Anderson et al. 1947; Libby et al. 1949; Arnold and Libby 1949) led to him being awarded the Nobel Prize for chemistry in 1960. The discovery of the method was followed by several years of work by a number of other people who noted that the initial assumption of constant atmospheric radiocarbon (^{14}C) levels seemed to be erroneous and, hence, that there is a need for calibration of radiocarbon ages onto the calendar scale (e.g. de Vries, 1958 and Willis et al., 1960).

Pivotal early work on estimating $p(\mu(\cdot)|\mathcal{X}, \mathcal{T})$ was undertaken by the influential physical chemist Hans Suess (1909 to 1993) who radiocarbon-dated samples of trees that had already been calendar-dated using tree-ring dating (dendrochronology). Plotting the resulting paired points, he hand-drew a "calibration" curve using a method that he called "cosmic schwung" (Suess 1967, 1970). Since then there have been continuous attempts to provide more reliable estimates of the relationship between radiocarbon and calendar age back over the entire period for which radiocarbon dating is viable (i.e. to about 55,000 or 60,000 years ago).

What has been clear since the very early work, is that the relationship is non-monotonic and so a single radiocarbon age often maps to multiple calendar ages. The non-monotonicity of this curve is one of the reasons that statisticians have, for so long, regarded radiocarbon dating as an interesting source of case studies. Indeed a statistician, Malcolm Clark, was highly influential in some of the early work on radiocarbon calibration curve estimation. He used the curve estimation problem as a case study in non-parametric curve fitting (Renfrew and Clark 1973, 1974) and, on at least one occasion (Clark 1975), gave considerable, tailored advice to radiocarbon scientists about error estimation. More recently, the IntCal Working Group (IWG) has sought to provide internationally-agreed estimates of calibration curves. Until recently, this group did not include a statistician and they used rather *ad hoc* procedures that involved taking weighted averages of calibration data in predefined windows and then using piece-wise linear or spline fits to create the curves (Pearson and Stuiver 1986; Stuiver and Pearson 1986a; Pearson and Stuiver 1993; Stuiver and Pearson 1993; Stuiver et al. 1998).

Once radiocarbon calibration curve estimates became routinely available, archaeologists and other researchers began to seek ways to obtain $p(\theta_0|\mu(\cdot), X_0)$. Since the radiocarbon dating laboratories provide only estimates of X_0 (of the form $x_0 \pm \sigma_0$) and we do not know $\mu(\cdot)$ precisely, this is clearly a statistical problem. The pioneering work in developing calibration algorithms and software was undertaken by radiocarbon scien-

tists (e.g. Stuiver and Reimer, 1986), but quite quickly statisticians started to see this as an interesting area of work. By 1989, when sufficient work had been done in this area to merit a review article (Aitchison et al. 1989), several statisticians were involved; in particular Tom Aitchison, Marian Scott and Morven Leese.

At that time there were, broadly speaking, two methods being used for calibration. The "intercept method" was popular in the early days because it could be implemented by anyone with access to a plot of the calibration curve at an appropriate scale. It simply involved drawing lines from the radiocarbon axis to intercept with the curve and then reading off the related calendar age(s) (Stuiver and Pearson 1986b). Since lines could be drawn at any point, both point estimates and estimates of confidence intervals were provided using this method. The alternative method, known as the "probability distribution method" was practical only for those with access to computing resources. Stuiver and Reimer (1989) discuss a range of options for estimating full probability distributions of calendar ages given the non-monotonic calibration curve.

At about this time, Bayesian statisticians were beginning to join the debate. Indeed even before the Aitchison et al. (1989) review was published, the pioneering paper which laid the groundwork for future research in Bayesian radiocarbon calibration had already been published. John Naylor and Adrian Smith were introduced to the interesting statistical challenges posed by radiocarbon calibration by Clive Orton, a statistician with considerable archaeological experience who could see the merits of formally utilising prior information in radiocarbon calibration. The paper that arose from this (Naylor and Smith 1988) is very detailed, offering a fully Bayesian modelling framework for radiocarbon calibration, not just for single radiocarbon determinations but for groups of related phases and sequences of determinations from a single archaeological site.

Within the radiocarbon literature debates about the right way to calibrate continued, but convergence towards the adoption of a generally Bayesian approach was happening relatively quickly. It took another statistician, Herold Dehling in collaboration with an influential radiocarbon scientist, Hans van der Plicht, to explain clearly to the radiocarbon community what formal options they had for calibration and why the Bayesian approach (formulated in the way advocated by Naylor and Smith, 1988) was superior to other approaches suggested in the past (Dehling and van der Plicht 1993).

Meanwhile, major developments were occurring in the Bayesian statistics research community. Adrian Smith, in collaboration with Alan Gelfand, following earlier work in this area by others including Geman and Geman (1984), drew attention to a number of simulation tools (stochastic substitution, Gibbs sampling and sampling-importance-resampling) for calculating numerical estimates of marginal posterior probability distributions (Gelfand and Smith 1990). With this work the so called MCMC (Markov chain Monte Carlo) "revolution" in Bayesian inference began.

By chance, at this time, one of us (CEB) took up a post as a research assistant on a project led by Adrian Smith and Cliff Litton. Their work involved developing and publishing case studies in the use of MCMC for Bayesian inference in complex real world problems. Given CEB's background in archaeological science, this early collab-

oration led to considerable investment of time and resources to provide archaeologists and other researchers with the tools that they needed to do fully Bayesian radiocarbon calibration (Buck et al. 1991, 1992, 1994a,c,b; Buck and Litton 1995; Zeidler et al. 1998; Blackwell and Buck 2003).

Thus, radiocarbon dating was a research problem in the right place at the right time. It was one of the very first complex statistical problems to benefit from the application of Bayesian inference via MCMC. For this reason it is one of the major case studies in the pivotal book *MCMC in Practice* (Gilks et al. 1996) and two Bayesian radiocarbon calibration packages (BCal and OxCal) are now routinely used by archaeologists in their research (Buck et al. 1999; Bronk Ramsey 1995, 2001). Indeed, the use of Bayesian statistics in radiocarbon calibration is now so well established that it was recently used as a case study by a philosopher of science. Daniel Steel used some of the work cited above, in an on-going debate in the philosophy of science about whether or not scientists reason partially, and therefore operate, as Bayesians (Steel 2001).

So why is it, more than ten years after the first routine use of Bayesian methods in radiocarbon calibration, that this topic is still worthy of being treated as a major case study in Bayesian statistics? The answer is simple. Until relatively recently Bayesian statisticians paid attention only to developing methods for estimating $p(\theta_0|\mu(\cdot), X_0)$; assuming that the calibration function, $\mu(\cdot)$, is well understood. In practice, however, this is far from the case and since 2002, when CEB joined the IntCal Working Group (IWG), which provided the 2004 internationally-agreed estimates of the radiocarbon calibrations curves (Reimer et al. 2004; Hughen et al. 2004; McCormac et al. 2004) substantial efforts have been made to bring modern Bayesian methods to bear upon upon inference for curve estimation as well. It is this work that forms the focus of what follows.

3 The calibration data

Before we can talk in detail about Bayesian methods for calibration curve estimation, we need to be clear what form the available calibration data take. As mentioned above, early curve estimates relied on radiocarbon dating lots of samples of trees that had previously been calendar dated via dendrochronology. Since tree-ring dating is assumed to give calendar ages that are reliable to the nearest year (and a calibration curve to the nearest year was considered more than adequate), researchers assumed that the only uncertainty associated with such data were the laboratory's estimates of their in-house errors on the radiocarbon measurements. Treating calendar age as the explanatory variable and radiocarbon age as the dependent variable meant that lots of standard curve fitting methods could be used. Unfortunately, however, this summary of the nature of the data belies an important complication that is present in most such data.

In practice, although tree-ring dating does provide a precise calendar date for each ring in each tree, there is not enough ^{14}C in a single ring to allow a reliable estimate of the radiocarbon age to be established by conventional (as opposed to atomic mass spectroscopy) radiocarbon dating laboratories. Consequently, for most calibration curve projects blocks of timber containing 10 or 20 rings are sent for radiocarbon dating. It is

this blocked nature of the data which led to the use of an *ad hoc* window-based method to calculate weighted averages for the 1993 and 1998 internationally-agreed radiocarbon calibration curves. As we will see, in 2004 we took account of this structure by formal modelling.

Recently a desire to provide calibration data back beyond the end of what is practical for tree-ring dating (currently around 12,000 years before present) has led to a search for other materials that can provide both calendar and radiocarbon date estimates. As a result, the database (available at http://www.radiocarbon.org/IntCal04.htm) used for the construction of the 2004 IntCal curve contains data derived from trees, but also from corals and from sequentially deposited lake/ocean sediments. Calcite from corals can be radiocarbon dated and calendar age estimates are usually obtained by uranium-series dating; also known as uranium/thorium (or U/Th) dating—another radiometric dating method but one that does not require calibration. For sedimentary sequences, calendar date estimates are either obtained by counting sequentially deposited annual layers (laminae or varves)—in a similar way to counting of tree-rings— or by using an age/depth model where such laminae do not exist. Trapped carbonaceous material such as foraminifera (small marine creatures that build a delicate shell, called a 'test', from calcite), are then selected from some layers and sent for radiocarbon dating. One of the important data sets for IntCal04 (Hughen et al. 2004) comprises laminated sediments, but does not have laminae all the way up to the present day. As a result, we know how many years are in the sample, but not when each layer was deposited. Such sequences—known as "floating" sequences—need to be tied into another record for which the calendar ages are known. This process of tying records together is known as "sequence matching" or "wiggle matching" and brings with it extra uncertainty in addition to that arising from the counting of annual laminae.

Clearly, each of these methods for obtaining calendar ages results in quite different error structures, some of which are highly correlated, that need to be carefully modelled in any formal curve estimation process.

4 The prior for $\mu(\cdot)$

Our prior beliefs about $\mu(\cdot)$ are crucial to the whole process of calibration. The use of calibration data makes sense only because we believe that $\mu(\theta)$ does not change too quickly with θ. We expect that $\mu(\cdot)$ should change fairly slowly over time, representing gradual changes in radiocarbon levels in the atmosphere. In fact, what we refer to here as "radiocarbon determinations" are not 'raw' ratios of carbon isotopes, but have been transformed onto an approximate timescale, by convention, in a way that would produce true ages if the original assumption of constant atmospheric radiocarbon were correct. We therefore might expect $\mu(\cdot)$ to change by approximately one (radiocarbon) year per (calendar) year.

We can represent these beliefs using a Wiener process, with drift β and variance r^2

(per year), relating radiocarbon levels at ages θ and θ' by

$$\mu(\theta')|\mu(\theta) \sim N((\theta' - \theta)\beta, (\theta' - \theta)r^2),\ \theta' > \theta.$$

In practice, since we wish to avoid the level of detailed modelling necessary to allow for intra-annual effects, we can use the natural annual time scale and think of $\mu(.)$ at integer ages only, with a prior given by a random walk with increments

$$Z_\theta \sim N(\beta, r^2).$$

5 Pointwise estimation of the calibration curve

5.1 Introduction

In this part of the paper, we describe the method actually used to estimate $\mu(\cdot)$ for the 2004 internationally-agreed Northern hemisphere terrestrial calibration curve (Buck and Blackwell 2004; Reimer et al. 2004). Extensions to this method (which are not discussed further here, but are documented in Buck and Blackwell, 2004) were used to provide the internationally-agreed marine calibration curve (Hughen et al. 2004) and the internationally-agreed Southern Hemisphere terrestrial calibration curve (McCormac et al. 2004). The development of the models and methods, and the actual calculations required to provide the 2004 estimates, were undertaken by the present authors, closely guided by the IntCal Working Group (IWG) of which CEB was a member.

At a meeting of the IWG at Queen's University Belfast (in 2002) criteria were agreed for the inclusion of calibration data which led to acceptance of datasets with the properties described in Section 3 (and documented in greater detail in Reimer et al., 2002). At a second meeting at the Woods Hole Oceanographic Institute (in 2003) the IWG also laid down some requirements for the statistical approach. In particular, having been encouraged by CEB and others at the Belfast meeting to quantify uncertainty on the calendar ages of the calibration data, the group was very keen that these should—for the first time—be taken into account. The IWG, which is composed mostly of laboratory physicists, also felt quite strongly that the statistical procedures should use an explicit repeatable numerical method (as opposed to a simulation-based approach). Finally, since it was not in the remit of the IWG to provide new software for radiocarbon calibration, it was also agreed that the final curves should be reported in the same form as previous ones; i.e. as posterior means and standard deviations for points $\mu(\theta)$ on the curve at selected regularly spaced values of θ.

The requirement of a numerical approach imposed some limits on what could be achieved on a timescale that was judged worthwhile, given the pressure from the radiocarbon dating user community for a curve that made use of the latest data. In particular, it was agreed, early on, that in representing prior beliefs of the form described in Section 4, it was not feasible to treat r as unknown, with uncertainty represented through a hyper-prior; instead, r needed to be fixed. Details of how a value was chosen are given in Section 5.5.1.

On the other hand, the decision to limit the output to means and variances meant that our calculations could treat separately each required point on the curve. This had the practical advantage that we could limit the observations to be considered in any one calculation to a fraction of the whole data set. This involves an element of approximation, but for each point on the curve we chose a suitable window so that this effect was negligible; see Section 5.5.2 for details. Of course, calculating points on the calibration curve separately has the disadvantage that it does not give the posterior covariances between the estimated points. However, given the way in which the curves are routinely used for calibration, it was judged that these covariances were unlikely to be used even if reported. This issue is revisited in Section 6.6.

For each calendar date θ^* at which we wish to estimate the curve, we choose a subset of the observations, $X_1, \ldots, X_N$, from dates in a suitable window around θ^*. As shown below, we can then approximate the joint distribution of $X_1, \ldots, X_N$ given $\mu(\theta^*)$ by a multivariate normal distribution, and invert this numerically to get a posterior mean and standard deviation for $\mu(\theta^*)$. This gives the required information at each point on the curve.

Specific models for $X_j|\mu(\theta^*)$ corresponding to the various kinds of calibration data are discussed in the following Sections 5.2 to 5.4. Some practical details of the implementation are discussed in Section 5.5, and the results described briefly in Section 5.6.

5.2 Modelling the calibration data - the simplest case

A standard assumption is that a radiocarbon determination, X_j, is an observation of the 'true' radiocarbon level of object j with an error, ϵ_j, made up of the laboratory's estimate of its internal error with variance, s_j^2 say, and what is known as a laboratory error multiplier, k_l say for laboratory l, which allows for inter-laboratory variability and for IntCal04 was explained in Reimer et al. (2004), pp. 1034–1035. In the simplest case, the calibration data come from samples of genuinely known date, so that $T_j = \theta_j, j = 1, \ldots, N$; conditional on $\mu(\cdot)$, the X_js are independent, with

$$\begin{aligned} X_j &= \mu(\theta_j) + \epsilon_j, \quad j = 1, \ldots, N \\ \epsilon_j &\sim N(0, \sigma_j^2), \quad j = 1, \ldots, N \end{aligned}$$

where $\sigma_j^2 = k_{l_j}^2 s_j^2$ and laboratory l_j is the source of the jth radiocarbon determination. If we wish to evaluate the curve at θ^* say, then conditional only on $\mu(\theta^*)$ we can write

$$X_j \sim N(\mu(\theta^*) + (\theta_j - \theta^*)\beta,\ |\theta_j - \theta^*|r^2 + \sigma_j^2). \tag{2}$$

Note that the X_js are *not* independent conditional on $\mu(\theta^*)$ since the uncertainty in $\mu(\cdot)$ is common to multiple observations. The covariance between two observations X_i and X_j, conditional on $\mu(\theta^*)$, is

$$\mathrm{Cov}(X_i, X_j) = r^2(\min\{\theta_i, \theta_j\} - \theta^*) \tag{3}$$

if $\theta_i, \theta_j > \theta^*$, with a similar expression if $\theta_i, \theta_j < \theta^*$ and zero covariance otherwise.

In practice, for each date θ^* at which we wish to estimate the curve, we select those observations $X_1, \ldots, X_N$ associated with calendar dates in a suitable window around θ^*. Then $X_1, \ldots, X_N$ given $\mu(\theta^*)$ have a multivariate normal distribution with means, variances and covariances as given in (2) and (3). With a uniform (or indeed any Gaussian) prior for $\mu(\theta^*)$, it is then straightforward to obtain a posterior estimate for $\mu(\theta^*)$ and an associated standard deviation (inverting the covariance matrix numerically).

In this simplest case, the approach described here is related to the approaches of Christen and Nicholls (2000), Goméz Portugal Aguilar et al. (2002) and Buck et al. (2006). However, if we are to allow for the real complexity of uncertainty and dependence in the calibration data, as described in Section 3, then we need to represent a more general structure in the data.

5.3 Averaging over blocks

A single determination X_j may arise not from material from a single precise date, but from an average over successive years e.g. blocks of consecutive tree-rings. Extending the calculations to this case is straightforward in principle. In practice, we carry out the calculation by exploiting the natural annual time scale, since each X_j is a linear combination of $\mu(\theta)$s, and any change in $\mu(\cdot)$ is a sum of increments Z_θ indexed by (integer) date θ. As before, we use the X_js associated with calendar dates in some 'window' $(\theta_{\min}, \theta_{\max})$ around θ^*. We write

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix}, \quad \boldsymbol{M} = \begin{pmatrix} \mu(\theta_{\min}) \\ \vdots \\ \mu(\theta^*) \\ \vdots \\ \mu(\theta_{\max}) \end{pmatrix}, \quad \boldsymbol{Z} = \begin{pmatrix} -Z_{\theta_{\min}+1} \\ \vdots \\ -Z_{\theta^*} \\ Z_{\theta^*+1} \\ \vdots \\ Z_{\theta_{\max}} \end{pmatrix}.$$

Then conditional on $\mu(\cdot)$,

$$\boldsymbol{X} = A\boldsymbol{M} + \boldsymbol{\varepsilon},$$

where the matrix A defines the contributions of different years to each sample. In practice, we typically have equal weights over n_j successive years for the jth sample, so $a_{j,\theta} = n_j^{-1}$ if $\theta_j \leq \theta \leq \theta_j + n_j - 1$ and $a_{j,\theta} = 0$ otherwise. If we want to condition solely on $\mu(\theta^*)$, we can write

$$\boldsymbol{M} = \mu(\theta^*)\mathbf{1} + B\boldsymbol{Z}$$

where the matrix B is of the form

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix},$$

extended to cover the range $(\theta_{\min}, \theta_{\max})$, and so

$$\boldsymbol{X} = \mu(\theta^*)\mathbf{1} + AB\boldsymbol{Z} + \boldsymbol{\varepsilon}$$

since each row of A sums to 1. This enables variances and covariances to be calculated readily.

The approach described so far enables us to estimate individual points on the calibration curve at dates where error on the calendar scale is negligible—essentially those using tree-ring data only. In particular, this approach suffices for the most recent parts of IntCal04 (Reimer et al. 2004), back to 12,400 years before present.

5.4 Uncertainty in calendar dates

5.4.1 Means and variances

Often in practice (for example when U/Th dating is used) we do not observe θ_j directly, but instead observe

$$T_j \sim N(\theta_j, \tau_j^2).$$

Given a flat, improper prior distribution for θ_j, this implies

$$\theta_j | T_j \sim N(T_j, \tau_j^2).$$

Then conditional on T_j and $\mu(\theta^*)$, X_j has mean

$$\mu(\theta^*) + (T_j - \theta^*)\beta$$

and variance

$$r^2 \mathrm{E}[|\theta_j - \theta^*| \,|\, T_j] + \sigma_j^2,$$

which can be calculated analytically when $\theta_j \sim N(T_j, \tau_j^2)$.

5.4.2 Covariances: independent calendar ages

The corresponding covariance between observations X_i and X_j is straightforward if we make the assumptions that T_i and T_j are independent and that θ_i and θ_j are independent, *a priori* and hence conditional on T_i and T_j. This is natural if for example the observations relate to completely separate data sources, or are directly dated using U/Th dating.

In such cases, $\mathrm{Cov}(X_i, X_j)$ involves

$$\mathrm{E}[c(\theta_i - \theta^*, \theta_j - \theta^*) | T_i, T_j]$$

where

$$c(\phi, \psi) = \begin{cases} \min\{|\phi|, |\psi|\} & \text{if } \phi, \psi \text{ have the same sign,} \\ 0 & \text{otherwise.} \end{cases}$$

This expectation cannot be evaluated analytically, but can be obtained readily by numerical integration.

5.4.3 Covariances: dependent calendar ages

The corresponding covariance between observations X_i and X_j is more complicated if the uncertainties in θ_i and θ_j are linked, for example if they have some layer-counting or sequence matching error in common.

For example, if a set of observations $\mathcal{J}$ belongs to the same 'floating' sequence of observations, assigned absolute dates by wiggle-matching, then any wiggle-matching error, ρ, will be common to them all, and so for each $j \in \mathcal{J}$ we might have

$$\begin{aligned} T_j &= \theta_j + \delta_j + \omega, \\ \delta_j &\sim N(0, \tau_j^2), \\ \omega &\sim N(0, \rho^2), \end{aligned}$$

with $\tau_j = 0$ if there is no calendar age error on θ_j other than that from wiggle-matching.

If, on the other hand, T_i and T_j are determined by layer counting within the same varved sequence, and $T_i < T_j$, then any error accrued up to layer T_i will also affect T_j, and we might have

$$\begin{aligned} T_i &= \theta_i + \omega_i, \\ T_j &= \theta_j + \omega_i + \omega_j \end{aligned}$$

where both ω_i and ω_j have Poisson distributions that, like the radiocarbon counting errors, can be approximated by normal distributions for present purposes.

In general, we assume that

$$\theta_i, \theta_j | T_i, T_j$$

is bivariate normal, as above, and then numerical integration can still be used to evaluate $\mathrm{Cov}(X_i, X_j)$.

A final example of dependence structures that we need to represent arises when, as well as having calendar date estimates, we know *a priori* that certain observations are ordered in time. For example, the sedimentary sequences described in Section 3 have annual layers ('laminae') whose ordering is known and should be incorporated in the analysis. Representing a strict ordering is not easy within the current framework, although it is straightforward in the more flexible approach described in Section 6. The estimated calendar dates of these layers respect the ordering, of course; for the moment, we simply add a strong positive correlation between the errors on the dates, as an approximate representation of the ordering.

The cases described in this section complete the range of error structures needed to permit a coherent, if in places approximate, representation of all the data available for the estimation of the older part of IntCal04 (Reimer et al. 2004) from 12,400 to 26,000 years before present.

5.5 Implementation for IntCal04

5.5.1 Specifying the random walk

The implementation outlined in Section 5.1 depends crucially on specifying the prior for μ, as described in Section 4, in terms of fixed numerical values for the mean β and variance (per year) r^2.

As indicated in Section 4, it seems natural to take $\beta = 1$, since that corresponds to the calibration curve changing by approximately 1 (radiocarbon) year per (calendar) year. In practice, however, the results are very robust to changes in β; even taking $\beta = 0$, which simplifies the calculations slightly, makes a negligible difference to the curve that we obtain.

In contrast, the value of the variance r^2 can make a real difference to the curve, so care is needed in setting this parameter. Christen and Nicholls (2000) estimated a value of $r = 7.9$ using single year data from Stuiver et al. (1998). In a similar way, we used high-resolution subsets of the IntCal04 data to estimate r. In the light of both sets of experiments, for the actual production of the curve in Reimer et al. (2004) we took $r = 8$.

5.5.2 Choosing the window

Care is needed in choosing the window around a given date θ^* for selecting the data to be used to calculate $\mu(\theta^*)$. Since the density of data points varies along the curve, the most natural approach is to use those observations X_j with values T_j, or θ_j if known, closest to θ^*. Experimentation showed that a surprisingly large number of data points had appreciable influence on any given $\mu(\theta^*)$. We also investigated the use of more elaborate rules for selecting the data points, based on their uncertainties, but this did not reduce the number needed to any great extent. In practice, we simply used the points with T_j or θ_j closest to θ^*, and increased the number chosen until the calculated values were essentially unchanged by adding further data. Along most of the curve, it was sufficient to use the 100 closest observations; for a small part where there were many observations with large errors, it was necessary to use the nearest 200 points to estimate $\mu(\theta^*)$ accurately.

5.5.3 Choosing output values

Since the curve was to be reported as a discrete set of points, we needed to choose an actual set of values of θ^* to use. For the most recent part of the curve, we chose to report values at 5 year intervals, which seemed to be fine enough to represent essentially all of the information in the data and which was compatible with the 10 year intervals with which users are familiar from IntCal98. Older parts of the curve are reported more sparsely, in view of the decreased density of the data and the increased calendar-age error.

It is worth emphasising that using the approach described, the curve can be evaluated

at *any* point θ^*, by contrast with earlier approaches where the spacing of points was built in to the method used.

5.5.4 Numerical integration

The calculation of covariances between observations, given calendar age error as in Section 5.4, requires evaluating the expectation of

$$c(\phi, \psi) = \begin{cases} \min\{|\phi|, |\psi|\} & \text{if } \phi, \psi \text{ have the same sign,} \\ 0 & \text{otherwise} \end{cases}$$

when ϕ, ψ have a bivariate normal distribution, by numerical integration. For IntCal, a rather crude method was used; evaluating $c(\cdot, \cdot)$ on a regular 25×25 grid, suitably centred, gave an adequate approximation.

5.6 The IntCal04 results

Figure 1 shows a section of IntCal04 (the 2004 internationally-agreed estimate of the Northern Hemisphere atmospheric radiocarbon calibration curve) which was published as Reimer et al. (2004). Posterior means for the levels of the curve at regular intervals are shown as a continuous curve, since that is effectively how they are treated when used (and provides clarity at this scale). An envelope based on one posterior standard deviation either side of the mean, again calculated pointwise, is shown in a similar way.

Clearly the posterior uncertainty on the curve varies a lot over the range of dates shown; obviously it is greater where the data points are sparse or uncertain (on either axis), but it also depends in a complex way on the particular configuration of the data points and the dependence between them. The smoothness of the posterior mean also varies with the data in a complex way. It is worth emphasising that the smoother parts of the posterior mean curve do not necessarily indicate that the calibration curve itself is believed to be smoother there; often it simply means that the precise shape of local fluctuations is not well determined by the data.

6 A Markov chain Monte Carlo approach

The requirements of the IWG, as described in Section 5.1, precluded simulation-based calculation of the curve at that time. However, during the process of modelling and calculation it became clear that there were a number of potential advantages of such an approach. In particular, a Markov chain Monte Carlo (MCMC) approach can incorporate uncertainty about the random walk parameters r^2 and β, and allows us to obtain posterior covariances between points on the curve, as well as estimates and standard deviations for any other aspect of the curve. We now describe how a curve can be constructed using MCMC, and briefly explore some consequences.

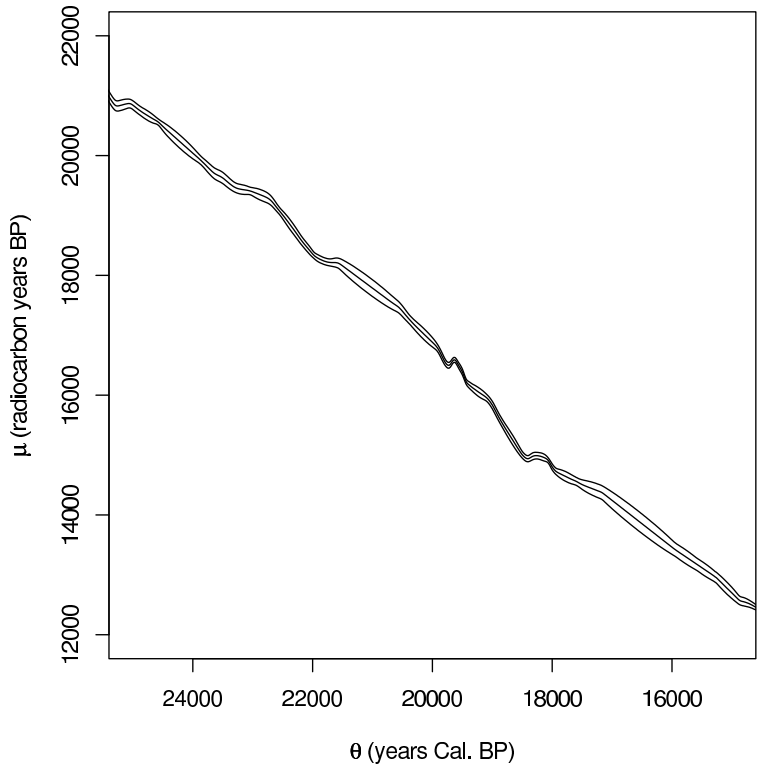

Figure 1: IntCal04 posterior mean curve with pointwise 1 S.D. envelope.

6.1 Preliminaries

Our approach will be to exploit the fact that the calibration data give direct information on the curve only at the particular, perhaps unknown, dates to which they relate. Thus

$$p(\mu|\mathcal{X},\mathcal{T}) = \int p(\mu|\mathcal{M},\Theta)p(\mathcal{M},\Theta|\mathcal{X},\mathcal{T})\mathrm{d}\mathcal{M}\mathrm{d}\Theta. \tag{4}$$

A crucial standard result for both terms in the integral in equation 4, which follows from modelling the curve as a Wiener process, is that the conditional distribution of $\mu(\theta^*)$ say is

$$\mu(\theta^*)|\mathcal{M},\Theta,r,\beta \sim N(\mu^B + (\mu^A - \mu^B)\phi, r^2(\theta^A - \theta^B)\phi(1-\phi)) \tag{5}$$

where

$$\begin{aligned} \theta^A &= \min\{\theta \in \Theta : \theta \geq \theta^*\} \\ \theta^B &= \max\{\theta \in \Theta : \theta \leq \theta^*\} \\ \mu^A &= \mu(\theta^A) \\ \mu^B &= \mu(\theta^B) \\ \phi &= \frac{\theta^* - \theta^B}{\theta^A - \theta^B}. \end{aligned}$$

To make inference about $\mathcal{M}$ based on observations $\mathcal{X} = \{X_1, \ldots, X_N\}$, we have

$$\mu_j | \mathcal{M}_{(j)}, \Theta, \mathcal{X}, r, \beta \sim N\left(\frac{x_j/\sigma_j^2 + \mu_{RW}/\sigma_{RW}^2}{1/\sigma_j^2 + 1/\sigma_{RW}^2}, \frac{1}{1/\sigma_j^2 + 1/\sigma_{RW}^2}\right) \tag{6}$$

where we write $\mathcal{M}_{(j)}$ for the vector with elements $\mu_i, i \neq j$ (and similarly for the θs, when necessary) and where μ_{RW} and σ_{RW}^2 are the mean and variance on the right hand side of equation 5 when used to calculate $\mu_j | \mathcal{M}_{(j)}, \Theta, r, \beta$.

Strictly speaking, the above formulation applies only to data without 'blocking' in the sense of Section 5.3. If there is blocking, then an observation tells us directly—although not separately—about a consecutive sequence of θs. This is straightforward in principle, but requires a slight extension of notation which we omit here for simplicity.

6.2 Known-age calibration samples

The results in Section 6.1 are enough to allow inference about the curve in the absence of calendar age errors, the equivalent of Section 5.2. Equation 6 can be used for a Gibbs sampling approach to imputing $\mathcal{M}$, since it gives the full conditional distributions $\mu_j | \mathcal{M}_{(j)}, \Theta, r, \beta$.

Again, we do not cover here the case of 'blocking' of observations. In principle, the updating for that case is straightforward but the coding of it is slightly intricate because of the need to cover various cases of overlapping blocks in generalising equation 5.

It is straightforward also to include Gibbs sampling steps for r and β, allowing the more general form of prior in which r and β are treated as unknown.

6.3 Independent calendar age errors

If our observations have uncertainty on the time axis, we use a Metropolis-within-Gibbs algorithm that updates each (θ_j, μ_j) together. First, write

$$\begin{aligned} & p(\theta_j, \mu_j | \mathcal{X}, \mathcal{T}, \mathcal{M}_{(j)}, \Theta_{(j)}, r, \beta) \\ = \; & p(\mu_j | \theta_j, x_j, \mathcal{M}_{(j)}, \Theta_{(j)}, r, \beta) p(\theta_j | x_j, t_j, \mathcal{M}_{(j)}, \Theta_{(j)}, r, \beta) \\ \propto \; & p(\mu_j | \theta_j, x_j, \mathcal{M}_{(j)}, \Theta_{(j)}, r, \beta) p(x_j | \theta_j, \mathcal{M}_{(j)}, \Theta_{(j)}, r, \beta) p(\theta_j | t_j). \end{aligned}$$

Various types of proposal distribution are possible; our choice exploits the analytic results above. We take

$$q(\theta'_j, \mu'_j|\theta_j, \mu_j) \propto g(\theta'_j - \theta_j)p(\mu'_j|\theta'_j, x_j, \mathcal{M}_{(j)}, \Theta_{(j)}, r, \beta)$$

for some symmetric density $g(\cdot)$; in practice we can use a suitably scaled Gaussian. Then the Hastings ratio is

$$\begin{aligned} & \frac{p(\theta'_j, \mu'_j|\mathcal{X}, \mathcal{T}, \mathcal{M}_{(j)}, \Theta_{(j)}, r, \beta)q(\theta_j, \mu_j|\theta'_j, \mu'_j)}{p(\theta_j, \mu_j|\mathcal{X}, \mathcal{T}, \mathcal{M}_{(j)}, \Theta_{(j)}, r, \beta)q(\theta'_j, \mu'_j|\theta_j, \mu_j)} \\ = & \frac{p(x_j|\theta'_j, \mathcal{M}_{(j)}, \Theta_{(j)}, r, \beta)p(t_j|\theta'_j)}{p(x_j|\theta_j, \mathcal{M}_{(j)}, \Theta_{(j)}, r, \beta)p(t_j|\theta_j)}, \end{aligned} \tag{7}$$

using the uniformity of the prior for θ_j. The algorithm can be summarised as follows. We propose a new value for θ_j from a symmetric distribution centred at the current value, and a new value for μ_j from the distribution in equation 6; acceptance is based on the likelihood for t_j and x_j, as shown in equation 7. The complex dependence of μ_j on Θ and $\mathcal{M}_{(j)}$ is mirrored in the proposal distribution, allowing acceptance to depend only on the likelihood.

6.4 Dependent calendar age errors

When we have dependence between the errors on the time axis, the algorithm of the previous section must be modified slightly. We can no longer make the simplification that $p(\theta_j|t_j, \mathcal{M}_{(j)}, \Theta_{(j)}, r, \beta) = p(\theta_j|t_j)$; instead we have $p(\theta_j|t_j, \mathcal{M}_{(j)}, \Theta_{(j)}, r, \beta) = p(\theta_j|t_j, \Theta_{(j)})$. The simplest change is therefore to replace the Hastings ratio of equation 7 with

$$\frac{p(x_j|\theta'_j, \mathcal{M}_{(j)}, \Theta_{(j)}, r, \beta)p(\theta'_j|t_j, \Theta_{(j)})}{p(x_j|\theta_j, \mathcal{M}_{(j)}, \Theta_{(j)}, r, \beta)p(\theta_j|t_j, \Theta_{(j)})}.$$

However, there is often a parameterisation that expresses the dependence in an interpretable way. In that case, it is convenient to include such parameters explicitly. For example, if there is an error term due to the dating of a 'floating' chronology (see Section 3), then that will be common to all observations within that chronology. We then have

$$T_j = \theta_j + \delta_j + \omega,$$

for all $j \in \mathcal{J}$ say, where $\delta_j \sim N(0, \tau_j)$ and $\omega \sim N(0, \rho)$ are all independent, as in Section 5.4. We can then choose to incorporate the 'shift' ω in our parameterisation. We can update each θ_j as before, with t_j replaced by $t_j - \omega$, and ω itself can be updated using a simple Gibbs step, since $p(\omega|\mathcal{X}, \mathcal{T}, \mathcal{M}, \Theta, r, \beta) = p(\omega|t_j - \theta_j : j \in \mathcal{J})$. This does not necessarily lead to the most efficient algorithm, but it is simple to implement and to interpret; this is the approach used in the example in Section 6.5.

6.5 Representing the curve

In this section, we consider briefly the use of the MCMC approach to reconstruct a part of the calibration curve. We limit ourselves to the interval 26000BP to 15000BP, to

avoid the complications due to blocking alluded to in Section 6.2. The aim is simply to illustrate the process; actual updating of the curve using this approach is work in progress.

The reconstruction is based on 232 data points, with varying amounts of calendar age uncertainty. Two MCMC runs, each of 20,000 iterations with a burn-in of 5000 iterations, were used. As indicated in Section 6.1, the 'curve' as it would be reported consists of a sample from the joint posterior distribution of $\mathcal{M}, \Theta$. A summary of this is shown in Figure 2, along with a representation of the relevant calibration data, and an extract is shown in more detail in Figure 3.

For each (θ_j, μ_j) an approximate 68% posterior region is given (to match the convention in the radiocarbon community of using one standard deviation for univariate uncertainty.) These regions are drawn in Figures 2 and 3 with solid lines (in blue, in the online/colour version). In practice, since the posterior distributions appear to be convex, the regions are obtained by peeling (Green 1980). This involves removing the 'outermost' points, defined to be those forming vertices of the convex hull of the posterior sample (i.e. the smallest convex set containing the sample), and then continuing the process, by removing the vertices of the convex hull of the remaining points, repeating until only the required proportion of points (here, 68%) remains.

The calibration data are represented by ellipses with 68% probability coverage; these can be thought of as normalised likelihoods, or equivalently as posteriors based on analysing each point separately, with uniform priors. These ellipses are drawn in Figures 2 and 3 with dashed lines (in red, in the online/colour version). The dotted lines in Figure 3 indicate the pairings between data and posteriors.

For example, the right-most region in Figure 3 corresponds to a data point with rather little calendar age (x-axis) uncertainty, but substantial radiocarbon level (y-axis) uncertainty. The next two overlapping regions correspond to two data-points with substantial uncertainty on both axes; the alignment of each region indicates the posterior dependence between θ_j and μ_j, while the overlap on the calendar age axis indicates that we cannot be sure which of these samples is actually the older. Figure 3 does not, of course, indicate the dependence between say (θ_i, μ_i) and (θ_j, μ_j), but that information *is* available in the MCMC output.

6.6 Calibration with MCMC

It is beyond the scope of this paper to describe in detail the various approaches that can be taken to calibration, given information on $\mu(\cdot)$, but it seems useful to give a brief outline of the two obvious methods. The more conservative would be to use the information on $(\mathcal{M}, \Theta)$ to generate a conventional calibration curve i.e. means and variances at regular intervals. More natural from the Bayesian perspective however is to calibrate a point X_0, say, directly from the posterior for $(\mathcal{M}, \Theta)$ by a further MCMC calculation, keeping $(\mathcal{M}, \Theta)$ fixed but generating (μ_0, θ_0) just as described in Section 6.3. The latter approach makes full use of the information available, including the posterior covariance between points on the curve.

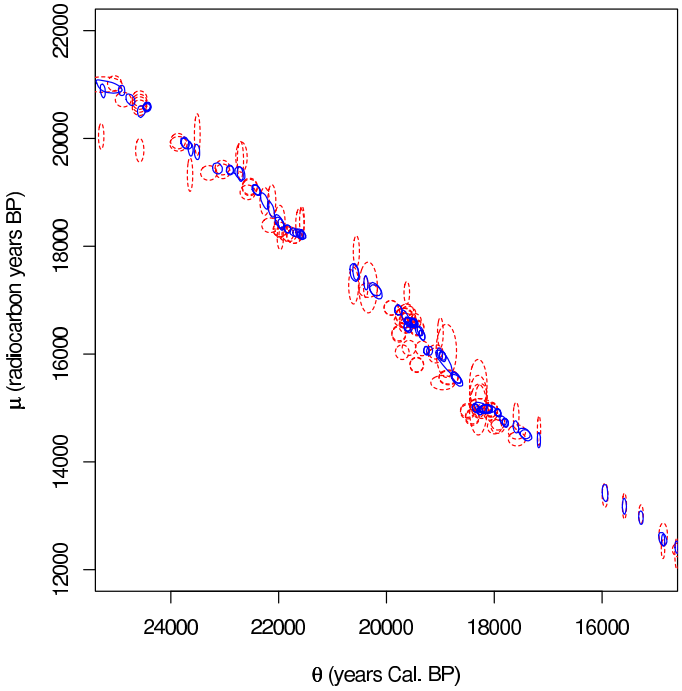

Figure 2: Data (dashed lines; red) and posterior summaries (solid lines; blue) for $(\Theta, \mathcal{M})$.

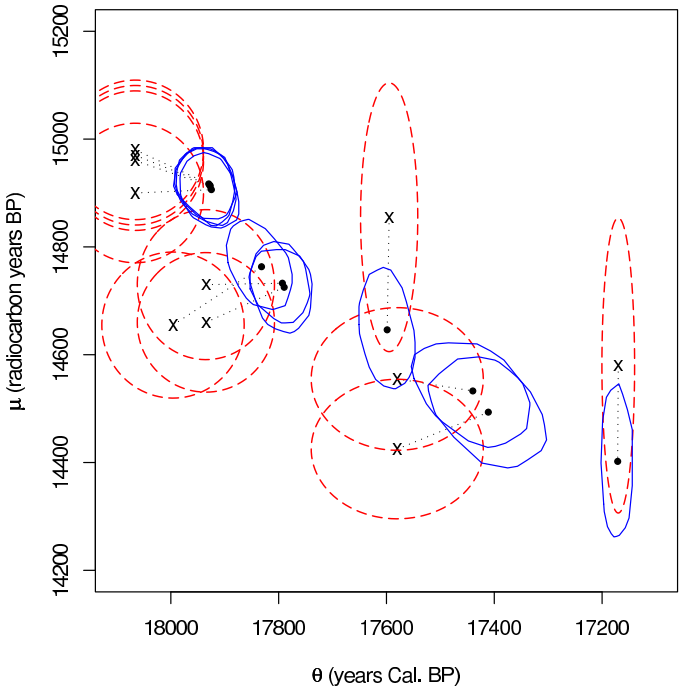

Figure 3: Selected data (dashed lines; red) and posterior summaries (solid lines; blue) for $(\Theta, \mathcal{M})$. Dotted lines link observations to corresponding posterior means.

One simple example shows the potential importance of the information available from the MCMC approach but not from conventional curve construction. Imagine that we wish to make inferences about the dates θ' and θ'' of *two* samples with radiocarbon determinations $X' = 14000 \pm 50$ and $X'' = 14100 \pm 50$. The marginal inferences are rather insensitive to the covariance between points on the calibration curve. However, if we are interested in say $\theta'' - \theta'$, the covariance is crucial. Figure 4 shows the posterior distributions for $\theta'' - \theta'$ with and without the covariance information; that is, firstly using the full information from the MCMC output, and secondly using a conventional approach to calibration that treats the reconstructed points on the calibration curve as independent. It is clear that including the covariance information substantially changes the interpretation of this result.

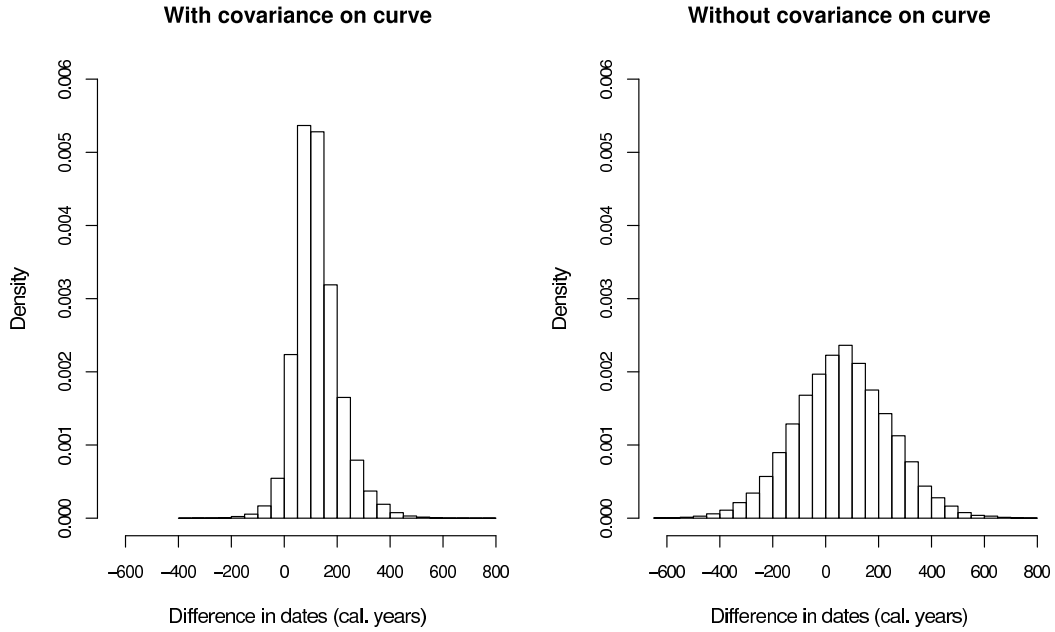

Figure 4: Posteriors for $\theta'' - \theta'$ when $X' = 14000 \pm 50, X'' = 14100 \pm 50$.

7 Discussion

By taking a Bayesian approach to radiocarbon calibration curve estimation, we have provided a methodology that takes account of sources of calendar age uncertainty that, for decades, had been ignored. We have also encouraged the radiocarbon dating community to see ^{14}C variation as a process about which we have at least some prior information that should be accounted for in our curve estimation methods. Alongside this, we have built a rapport with the IWG and are making plans to update the curve

estimates again. This on-going work will be considerably enhanced by a recent grant from the UK's Natural Environment Research Council (NERC) which will fund the provision of both MCMC methodology and user-friendly code that will be used for future updates of the calibration curves.

Other statistical aspects that we aim to investigate include the following.

- There are several aspects of calibration data collation and management that, until now, we have not had the resources to investigate, but would certainly benefit from input from statisticians. One example is that data providers currently each undertake their own "wiggle matching" of floating sequences. This has meant that each match has been undertaken independently of any others and that all have been undertaken separately from curve estimation. We would like to find a way for these to be done altogether.

- Many of the data points in the IntCal database (e.g. those from ocean cores) relate to samples from organisms that metabolised carbon from marine environments. For such data to be useful in estimating the terrestrial calibration curve, a location-specific marine/terrestrial offset is applied. The approach taken in 2004 implies that we have reliable knowledge of the modern offset and that the offset does not vary over time. There is increasing evidence that the latter assumption does not hold. Thus, as part of the new NERC funded project, we will be collecting data that will allow us to investigate the scale and nature of the variability in offset over time.

- There is prior information about $\mu(\cdot)$ that we are not currently using. For example, solar cycles are known to impact the production of ^{14}C and it would be interesting to investigate the impact of a component of periodic behaviour in our prior. This seems potentially important since there are cycles, with periods 11 and 25 years for example, detectable in our 2004 estimate of the curve.

- Given the results shown in Figure 4, we would like to begin to encourage the radiocarbon user community to see the calibration curve $\mu(\cdot)$ as more than simply a collection of independent points and standard deviations, and to make appropriate allowance for its dependence structure when using it for calibration.

- Taking the previous point a step further, we would like to continue our "knowledge transfer" work by conveying to the user community the idea that the estimation of $\mu(\cdot)$ and of θ_0 are not necessarily best seen as two separate steps. Almost all material sent for radiocarbon dating comes with some prior information about its calendar age, and sometimes this is quite informative. For example, some lake and ocean sediments do not have annual layers, but do provide ordering information and give rise to lots of radiocarbon determinations. Essentially, these are floating sequences for which we have poor prior knowledge of the calendar scale. We could use our model to combine such sequences with the data in the calibration database; giving us the potential to learn more about the calibration curve as well as about the chronology of the sediments. Since there are still several important

periods where the data in the calibration database are sparse, this may prove to be a very fruitful way forward. It may also prove controversial because users are currently discouraged from deriving their own calibration curve estimates, since this leads different research groups to obtain different calendar date estimates for the same samples. Our first step towards encouraging the radiocarbon research community to think more about this will be to demonstrate the impact of doing joint inference and to provide software to allow others to experiment.

References

Aitchison, T. C., Leese, M., J., D., Mook, W. G., Otlet, R. L., Ottaway, B. S., Pazdur, M. F., van der Plicht, J., Reimer, P. J., Robinson, S. W., Scott, E. M., Stuiver, M., and Weninger, B. (1989). "Comparison of methods used for the calibration of radiocarbon dates." *Radiocarbon*, 31(3): 846–864. 228

Aitken, M. J. (1990). *Science-based Dating in Archaeology*. Harlow, England: Longman Archaeology Series. 226

Anderson, E. C., Libby, W. F., Weinhouse, S., Reid, A. F., Kirshenbaum, A. D., and Grosse, A. V. (1947). "Natural Radiocarbon from Cosmic Radiation." *Physical Review*, 72(10): 931–936. 227

Arnold, J. R. and Libby, W. F. (1949). "Age Determinations by Radiocarbon Content: Checks with Samples of Known Age." *Science*, 110(2869): 678–680. 227

Blackwell, P. G. and Buck, C. E. (2003). "The Late Glacial human reoccupation of north western Europe: new approaches to space-time modelling." *Antiquity*, 77(296): 232–240. 229

Bowman, S. (1990). *Radiocarbon Dating*. London: British Museum Publications. 226

Bronk Ramsey, C. (1995). "Radiocarbon Calibration and Analysis of Stratigraphy: The OxCal Program." *Radiocarbon*, 37(2): 425–430. 229

— (2001). "Development of the Radiocarbon Program OxCal." *Radiocarbon*, 43(2A): 355–363. 229

Buck, C. E. and Blackwell, P. G. (2004). "Formal Statistical Models for Estimating Radiocarbon Calibration Curves." *Radiocarbon*, 46(3): 1093–1102. 231

Buck, C. E., Christen, J. A., and James, G. N. (1999). "BCal: an on-line Bayesian radiocarbon calibration tool." *Internet Archaeology*, 7. (http://intarch.ac.uk/journal/issue7/buck/). 229

Buck, C. E., Christen, J. A., Kenworthy, J. B., and Litton, C. D. (1994a). "Estimating the duration of archaeological activity using ^{14}C determinations." *Oxford Journal of Archaeology*, 13: 229–240. 229

Buck, C. E., Goméz Portugal Aguilar, D., Litton, C. D., , and O'Hagan, A. (2006). "Bayesian non-parametric estimation of the calibration curve for radiocarbon dating." *Bayesian Analysis*, 1(2): 265–288. http://ba.stat.cmu.edu/journal/2006/vol01/issue02/buck265-288.pdf. 233

Buck, C. E., Kenworthy, J. B., Litton, C. D., and Smith, A. F. M. (1991). "Combining archaeological and radiocarbon information: a Bayesian approach to calibration." *Antiquity*, 65: 808–821. 229

Buck, C. E. and Litton, C. D. (1995). "The radiocarbon chronology: further consideration of the Danebury dataset." In Cunliffe, B. (ed.), *Danebury: an iron age hillfort in Hampshire, Volume 6, A hillfort community in Hampshire*, 130–136. Council for British Archaeology, Report Number 102. 229

Buck, C. E., Litton, C. D., and Scott, E. M. (1994b). "Making the most of radiocarbon dating: some statistical considerations." *Antiquity*, 68: 252–263. 229

Buck, C. E., Litton, C. D., and Shennan, S. J. (1994c). "A case study in combining radiocarbon and archaeological information: the early Bronze Age settlement of St. Veit-Klinglberg, Land Salzburg, Austria." *Germania*, 72: 427–447. 229

Buck, C. E., Litton, C. D., and Smith, A. F. M. (1992). "Calibration of radiocarbon results pertaining to related archaeological events." *Journal of Archaeological Science*, 19: 497–512. 229

Christen, J. A. and Nicholls, G. (2000). "Random-walk radiocarbon calibration." Technical Report #457, Mathematics Department, University of Auckland. 233, 236

Clark, R. (1975). "A calibration curve for radiocarbon dates." *Antiquity*, 49: 251–266. 227

de Vries, H. (1958). "Variations in concentration of radiocarbon with time and location on earth." *Koninkl. Nederl. Akademie van Wetensch., Amsterdam, Proc. Ser. B.*, 61: 94–102. 227

Dehling, H. and van der Plicht, J. (1993). "Statistical Problems in Calibrating Radiocarbon-Dates." *Radiocarbon*, 35(1): 239–244. 228

Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling based approaches to calculating marginal densities." *Journal of the American Statistical Association*, 85: 398–409. 228

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions and Bayesian restoration of images." *I.E.E.E. Transactions of Pattern Analysis and Machine Intelligence*, 6: 721–741. 228

Gilks, W., Richardson, S., and Spiegelhalter, D. (eds.) (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall. 229

Goméz Portugal Aguilar, D., Litton, C. D., and O'Hagan, A. (2002). "Novel statistical model for a piece-wise linear radiocarbon calibration curve." *Radiocarbon*, 44(1): 195–212. 233

Green, P. J. (1980). "Peeling Bivariate Data." In Barnett, V. (ed.), *Interpreting Multivariate Data*, 3–19. Chichester: Wiley. 241

Hughen, K. A., Baillie, M. G. L., Bard, E., Beck, J. W., Bertrand, C. J. H., Blackwell, P. G., Buck, C. E., Burr, G. S., Cutler, K. B., Damon, P. E., Edwards, R. L., Fairbanks, R. G., Friedrich, M., Guilderson, T. P., Kromer, B., McCormac, G., Manning, S., Ramsey, C. B., Reimer, P. J., Reim er, R. W., Remmele, S., Southon, J. R., Stuiver, M., Talamo, S., Taylor, F. W., and van der Plicht, C. E., J. an d Weyhenmeyer (2004). "Marine04—marine radiocarbon age calibration, 0–26 cal kyr BP." *Radiocarbon*, 46(3): 1059–1086. 229, 230, 231

Libby, W. F., Anderson, E. C., and Arnold, J. R. (1949). "Age Determination by Radiocarbon Content: World-Wide Assay of Natural Radiocarbon." *Science*, 109(2827): 227–228. 227

McCormac, F. G., Hogg, A. G., Blackwell, P. G., Buck, C. E., Higham, T. F. G., and Reimer, P. J. (2004). "SHCal04—Southern Hemisphere calibration, 0–11.0 cal kyr BP." *Radiocarbon*, 46(3): 1087–1092. 229, 231

Naylor, J. C. and Smith, A. F. M. (1988). "An archaeological inference problem." *Journal of the American Statistical Association*, 83(403): 588 – 595. 228

Pearson, G. W. and Stuiver, M. (1986). "High-precision calibration of the radiocarbon time scale, 500–2500 BC." *Radiocarbon*, 28(2B): 839–862. 227

— (1993). "High-precision bidecadal calibration of the radiocarbon time scale, 500–2500 BC." *Radiocarbon*, 35(1): 25–33. 227

Reimer, P. J., Baillie, M. G. L., Bard, E., Bayliss, A., Beck, J. W., Bertrand, C. J. H., Blackwell, P. G., Buck, C. E., Burr, G. S., Cutler, K. B., Damon, P. E., Edwards, R. L., Fairbanks, R. . G., Friedrich, M., Guilderson, T. P., Hogg, A. G., Hughen, K. A., Kromer, B., G., M., Manni ng, S., Ramsey, C. B., Reimer, R. W., Remmele, S., Southon, J. R., Stuiver, M., Talamo, S., Taylor, F. W., van der Plicht, J., and Weyhenmeyer, C. E. (2004). "IntCal04—terrestrial radiocarbon age calibration, 0–26 cal kyr BP." *Radiocarbon*, 46(3): 1029–1058. 229, 231, 232, 234, 235, 236, 237

Reimer, P. J., Hughen, K. A., Guilderson, T. P., McCormac, G., Baillie, M. G. L., Bard, E., Barratt, P., Beck, J. W., Buck, C. E., Damon, P. E., Friedrich, M., Kromer, B., Ramsey, C. B., Reimer, R. W., Remmele, S., Southon, J. R., Stuiver, M., and van der Plicht, J. (2002). "Preliminary Report of the First Workshop of the IntCal04 Radiocarbon Calibration/Comparison Working Group." *Radiocarbon*, 44(3): 653–661. 231

Renfrew, C. and Clark, R. M. (1973). "Tree-ring calibration of radiocarbon dates and the chronology of ancient Egypt." *Nature*, 243: 266–270. 227

— (1974). "Problems of the radiocarbon calendar and its calibration." *Archaeometry*, 16: 5–18. 227

Steel, D. (2001). "Bayesian Statistics in Radiocarbon Calibration." *Philosophy of Science*, 63(3): S153–S164. Supplement: Proceedings of the 2000 Biennial Meeting of the Philosophy of Science Association. Part I: Contributed Papers. 229

Stuiver, M. and Pearson, G. W. (1986a). "High-precision calibration of the radiocarbon time scale, AD 1950–500 BC." *Radiocarbon*, 28(2B): 805–838. 227

— (1986b). "High-precision calibration of the radiocarbon time scale, AD 1950- 500 BC." *Radiocarbon*, 28(2B): 805–838. 228

— (1993). "High-precision bidecadal calibration of the radiocarbon time scale, AD 1950–500 BC and 2500–6000 BC." *Radiocarbon*, 35(1): 1–23. 227

Stuiver, M. and Reimer, P. (1986). "A computer program for radiocarbon age calibration." *Radiocarbon*, 28(2B): 1022–1030. 228

— (1989). "Histograms obtained from computerized radiocarbonage calibration." *Radiocarbon*, 31(3): 817–823. 228

Stuiver, M., Reimer, P. J., Bard, E., Beck, J. W., Burr, G. S., Hughen, K. A., Kromer, B., McCormac, F. G., v. d. Plicht, J., and Spurk, M. (1998). "INTCAL98 Radiocarbon Age Calibration, 24,000 – 0 cal BP." *Radiocarbon,*, 40: 1041–1083. 227, 236

Suess, H. E. (1967). "Bristlecone pine calibration of the radiocarbon time scale from 4100 B.C. to 1500 B.C." In *Proc. Symp. on Radiocarbon Dating and Methods of Low-Level Counting, Monaco, I.A.E.A., Vienna*, 143–151. 227

— (1970). "Bristlecone pine calibration of the radiocarbon time scale from 5400 B.C. to the present." In Olsson, I. U. (ed.), *Radiocarbon Variations and Absolute Chronology*, 303–12. Council for British Archaeology, Report Number 102. 227

Willis, E. H., Tauber, H., and Munnich, K. O. (1960). "Variations in the atmospheric radiocarbon concentration over the past 1300 years." *Radiocarbon*, 2: 1–4. 227

Zeidler, J. A., Buck, C. E., and Litton, C. D. (1998). "The integration of archaeological phase information and radiocarbon results from the Jama River Valley, Ecuador: a Bayesian approach." *Latin American Antiquity*, 9(2): 135–159. 229

Acknowledgments

We are grateful to members of the IntCal Working Group and, especially, to Paula Reimer who patiently organises the constantly growing number of people involved in the radiocarbon calibration curve estimation work.

We are also grateful to Andrew Millard and Michael Goldstein for helpful comments at an early stage in this work, to Andrew Parnell and John Haslett for their comments and questions on an early draft of this paper, which have helped to improve its clarity, and to Brad Carlin and anonymous referees for useful comments on the submitted version.