Radiocarbon dating with temporal order constraints

Geoff Nicholls
Department of Mathematics
nicholls@math.auckland.ac.nz

Martin Jones
Department of Anthropology
martin@car.ant.auckland.ac.nz

Auckland University, Private Bag 92019,
Auckland, New Zealand

April 1998
revised September 2000

Bayesian methods are now widely used for analysing radiocarbon dates. We find that the non-informative priors in use in the literature generate a bias towards wider date ranges which does not in general reflect substantial prior knowledge. We recommend using a prior in which the distribution of the difference between the earliest and latest dates has a uniform distribution. We show how such priors are derived from a simple physical model of the deposition and observation process. We illustrate this in a case study, examining the effect that various priors have on the reconstructed dates. Bayes factors are used to help decide model choice problems.

KEYWORDS: radiocarbon dating, stratigraphy, Bayesian inference, Stein effect.

Technical Report #407, Mathematics Department, University of Auckland, 1998

1 Introduction

The interpretation of chronometric data is fundamental to archaeology. While individual dates are of some interest, archaeologists are most interested in temporal parameters of cultural processes inferred from a body of chronometric data. The temporal duration, or span, of activity represented by the date set is one such parameter. However, cultural processes may develop over time scales which are small compared to the dating accuracy. Moreover, as we will see, analysed date sets typically contain heterogeneous data types. Some statistical analysis is needed. A suitable framework will handle multiple chronometric data types and explicitly quantify the uncertainty in any inferred temporal parameters.

In this section a simplified observation model is used to illustrate certain problems not dealt with elsewhere. In Section 2 we repeat the analysis without those simplifying assumptions, show that the problems remain, and suggest a solution. In Section 3 we interpret our solution. In Section 4 we explain how our approach works for other models in the literature. In Section 5 we mention some unrelated problems we encountered whilst analysing radiocarbon age data.

Consider the data in Table 1. Seven specimens of charcoal are taken from six layers of earth, older layers being deeper. A radiocarbon date measurement is made for each specimen. For the moment we will suppose that the measured radiocarbon date for each specimen is a realisation of an independent Gaussian random variable, with mean equal to the unknown true date for that specimen and known standard deviation. (Note that this is not the standard radiocarbon observation model described in Section 2.1 and assumed throughout the remainder of this paper. We will clarify what we mean by a date in Section 2.1.) We wish to estimate the unknown true date associated each specimen, taking into account the age order constraint arising from stratigraphy. Another quantity of interest is the span of time from the bottom of the first stratum to the top of the last.

Buck, Kenworthy, Litton and Smith have described a Bayesian method for estimating dates from radiocarbon measurements (see Buck *et al.* 1991 and Buck *et al.* 1992). Stratigraphic constraints may be incorporated in the analysis. The authors give a straightforward inference scheme, based on MCMC simulation from the posterior. Christen 1994 gives an outlier analysis, within the same Bayesian/MCMC framework. Statistical packages OxCal, described in Ramsey 1995, and BCal, described in Buck *et al.* 1999, implement some of the methods presented in Buck *et al.* 1992. See Buck *et al.* 1996 and Litton and Buck 1996 for reviews of the field.

The reference priors described in Buck *et al.* 1996 (and below in Sections 2.3 and 4) assign a constant density to all states and are intended to be reasonably non-informative. However, when dates cover an interval of time which is not large compared to the error in the radiocarbon dating process, we find that such priors are undesirably informative, pushing the earliest and latest reconstructed dates towards unrepresentatively early and late values respectively. This feature is related to the Stein effect (see Stein 1956). The unconstrained maximum likelihood estimator (MLE) for the seven unknown parameters is just the vector of data values themselves. The span over observed dates will typically exceed the span present in the unknown true dates. If this is an unfamiliar idea, consider generating synthetic data from seven given dates spanning twenty five years, adding a Gaussian variate drawn from a density with a standard deviation equal fifty years. This spreading tendency is strong when the standard deviation in the observation model is of the order of the unknown true span in the specimen ages. Data sets from New Zealand, with its relatively short prehistory, are often sensitive to this effect. See Section 6 for further discussion of the Stein effect. In the

absence of constraints, the MLE is the posterior mean for the constant-density prior and so it is unsurprising that the spreading bias in the MLE is present in the posterior distribution also. In this paper we explain how we were able to correct for this effect, by returning to a physical model for the quantities parameterised in the prior. The correction would be straightforward in the simplified model we are discussing at this point. It is not so when order constraints apply to parameters, and the radiocarbon observation model of Section 2.1 is in use.

The unknown true specimen dates were realised through the action of complex processes of deposition, aging, archaeological recovery and sample selection. In Section 3 we calculate prior date distributions arising under a simple family of process models. If we select the prior density by choosing the simplest, physically plausible, process model we can find, the problems of prior bias disappear. Indeed, these process-models-of-choice have the property that the marginal prior density of the span is uniform (up to a weak function of the span variable). They turn out to be non-informative in this practically important sense. Physics-based modeling of deposition processes plays a similar role in Christen *et al.* 1995, where peat age and depth data are analysed.

Is there a process model which would lead to a prior density which is a constant, *ie* uniform over all states? If we had such a model, we could then give a "physical" interpretation of the assumptions built into the reference priors of Buck *et al.* 1996 and the earlier literature. We give the appropriate process model at the end of Section 3. With the benefit of this interpretation from a physical model, it is easy to understand why the apparently non-informative constant prior is informative in respect of span.

The constant prior density contains information which is not representative of the knowledge which is typically available to the archaeologist. Although this information is often overwhelmed by the data, it is easy to find datasets from New Zealand in which the effect is strong enough to be troublesome. We present, in Section 2, a small case study. Two priors are compared, a first prior model, assigning constant density to all states, and a second prior, derived from a process model. The posterior distributions of selected summary statistics differ markedly between the two models. We compute a Bayes factor for the two models. This provides an objective criterion for favouring the second model on this particular data set. Details of the estimator used to compute the Bayes factor are given in Appendix B.

When the number of radiocarbon dates is small, it is possible to carry out the sampling needed for the inference in Section 2 using a rejection algorithm. This is not practical when the number of dates is at all large. Indeed MCMC simulation of the posterior density can be quite challenging under certain process model priors. Simple Gibbs and random-walk Metropolis-Hastings samplers are certainly inadequate. Details of sampling algorithms used in Section 2 are given in Appendix A. The key elements of a Metropolis-Hastings sampler adequate for more substantial problems are outlined in Appendix C. The priors and simulation procedures described in this paper have been taken up in some of the widely used Bayesian calibration software online. At present these include the OxCal software available at `www.rlaha.ox.ac.uk` and our own implementation, DateLab, which can be down-loaded from `www.car.auckland.ac.nz`.

2 Case study

2.1 Observation Model

Consider a single vertical series of M abutting layers of earth. It is helpful to think of the earth and datable artifacts piling up vertically over time to form the layers. Suppose N_m radiocarbon age determinations are made in layer m, making $K = \sum N_m$ dates in all. For $n \in \{1, 2 \ldots N_m\}$ let $y_{m,n}$ denote the value of the n'th radiocarbon age measured in the m'th layer, reported with associated standard error $\sigma^{(Y)}_{m,n}$. For all scalar quantities $X_{m,n}$ let X denote the corresponding vector, so that $y = (y_{1,1}, \ldots, y_{M,N_M})$, *etc.* Let $\Theta_{m,n}$ denote the unknown true date at which object (m, n), which was once part of a living organism, ceased exchanging carbon with its environment (this definition is crude, but adequate for our purposes). In archaeological chronometry the units of age are sometimes taken as years before 1950. We do not use that convention in this paper. The quantity $\Theta_{m,n}$ is quite simply a calendar year AD or BC as appropriate. There is no year zero. A quantity measured in calendar years is referred to as a *calibrated date*. The measurement $y_{m,n}$ is called a *conventional radiocarbon age* (CRA).

Due to variations in the relative abundance of carbon isotopes in the earth's atmosphere over the years, a CRA is not simply an age in years (so again standard usage is misleading). It is related to a corresponding calendar date by a calibration function, μ, a part of which is shown in Figure 1. If a is a calendar date and b is the CRA for that date then $b = \mu(a)$

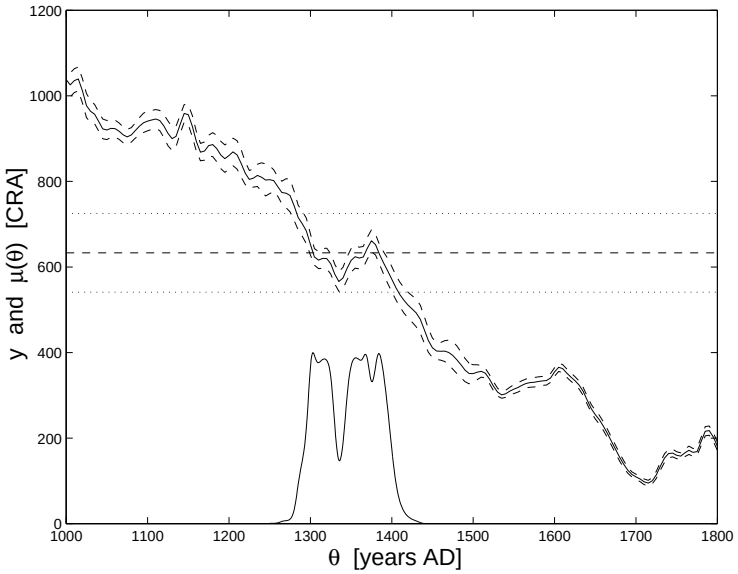

Figure 1: Calibration function $\mu(\theta)$ (solid) with $2\sigma^{(\mu)}(\theta)$ limits (dashed) for calibration of terrestrial (as opposed to marine) specimen, from Stuiver *et al.* 1998. Horizontal lines indicate a CRA (Conventional Radiocarbon Age determination, $y_{m,n}$, dashed) with $2\sigma^{(Y)}_{m,n}$ limits (dotted). The likelihood $\ell(y_{m,n}|\theta_{m,n}, \sigma)$ (solid) is drawn below the calibration curve.

in the absence of measurement noise. The calibration function μ and its standard error, $\sigma^{(\mu)}$, have been measured, and are available from `http://depts.washington.edu/qil/` at decadal intervals. We spline these decadal values to get functions μ and $\sigma^{(\mu)}$ which are

piecewise constant by year. When specimen (m, n) is of terrestrial origin, the terrestrial calibration function is used, otherwise a different marine calibration function is used. No marine materials are dated in our case study. A section of the terrestrial curve is shown in Figure 1. Because this curve is not monotone, a terrestrial CRA with no observation noise need not uniquely determine a calibrated date.

Following Buck *et al.* 1991, a standard observation model relating CRA's to calibrated dates is

$$y_{m,n} = \mu(\Theta_{m,n}) + \epsilon^{(Y)}_{m,n} + \epsilon^{(\mu)}_{m,n}|\Theta_{m,n} \tag{1}$$

where $\epsilon^{(Y)}_{m,n}$ and $\epsilon^{(\mu)}|\Theta_{m,n}$ are normal random variables drawn independently for each $m = 1, 2 \ldots M$ and each $n = 1, 2 \ldots N_m$ according to

$$\epsilon^{(Y)}_{m,n} \sim \mathsf{N}(0, {\sigma^{(Y)}_{m,n}}^2),$$

and

$$\epsilon^{(\mu)}_{m,n}|\Theta_{m,n} \sim \mathsf{N}(0, \sigma^{(\mu)}(\Theta_{m,n})^2).$$

Let $\theta_{m,n}$ be a trial value for $\Theta_{m,n}$ and let $\ell(y_{m,n}|\theta_{m,n})$ give the likelihood for an individual calibrated date. Let $\sigma(\Theta_{m,n})^2$ denote the net variance of a CRA,

$$\sigma(\Theta_{m,n})^2 = \sigma^{(\mu)}(\Theta_{m,n})^2 + {\sigma^{(Y)}_{m,n}}^2. \tag{2}$$

The likelihood

$$\ell(y_{m,n}|\theta_{m,n}) = \frac{\exp(-(y_{m,n} - \mu(\theta_{m,n}))^2/2\sigma(\theta_{m,n})^2)}{\sqrt{2\pi\sigma(\theta_{m,n})^2}} \tag{3}$$

is not a Gaussian in $\theta_{m,n}$. Note that we are conditioning on knowledge of the standard errors on the right in Equation (2) although these are subject to measurement error. Observations are assumed independent, so the joint likelihood, $L(y|\theta)$ say, is

$$L(y|\theta) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} \ell(y_{m,n}|\theta_{m,n}).$$

In Section 5 we return to the observation model in order to point out errors in the present treatment of reservoir offsets.

2.2 Data

Consider the data shown in Table 1, a set of seven CRA's from a single series of six layers. Likelihood curves for each date are displayed in Figure 2. These data are a subset of a large set of dates gathered at the mouth of Shag river, in southern New Zealand (Anderson *et al.* 1996). We have taken all the charcoal dates (*ie* the dated material was charcoal, as opposed to say eggshell, or shellfish) from that larger set. Quantities of interest to archaeologists include the length of time for which the site was occupied, and the actual dates at which occupation began and ended.

Outliers are a significant problem in the interpretation of radiocarbon dates. Visual inspection of graphs like Figure 2 can be helpful when stratigraphic constraints are relevant. See Christen 1994 for a discussion of this issue in a similar setting. Since the charcoal of short lived species yields relatively reliable CRA's, and Figure 2 shows no obvious problems, we will assume that our observation model is good for all our data.

Date	(m,n)	$y_{m,n}$	$\sigma^{(Y)}_{m,n}$
NZ 7758	$(1,1)$	580	47
NZ 7761	$(2,1)$	600	50
NZ 7757	$(3,1)$	537	44
NZ 7756	$(4,1)$	670	47
NZ 7755	$(5,1)$	646	47
WK 2589	$(5,2)$	630	35
NZ 7771	$(6,1)$	660	46

Table 1: The table shows all charcoal dates from the SM/C:Dune series obtained from Shag river mouth, New Zealand and reported in Anderson *et al.* 1996. The column $y_{m,n}$ lists CRA's, with no corrections. The column (m,n) lists m, the layer index and n, the specimen index within a layer.

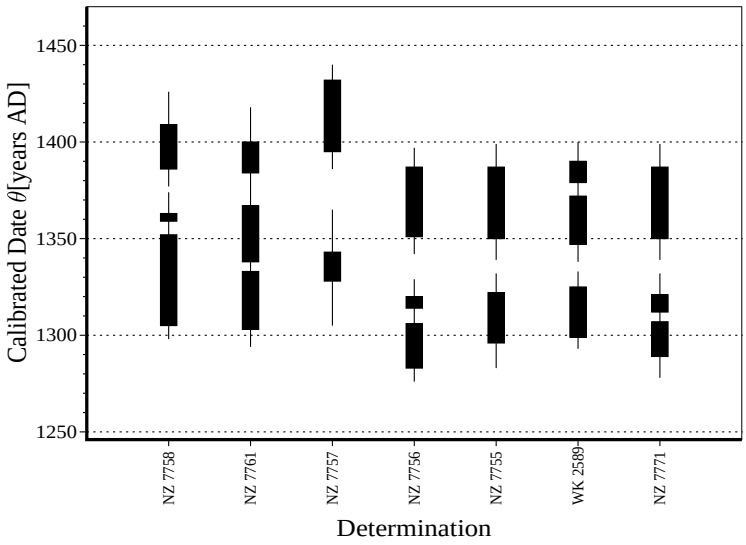

Figure 2: The graph represents the likelihoods for the data in Table 1, calculated using the observation model of Equation (1) and the calibration curves $\mu(\theta_{m,n})$ and $\sigma^{(\mu)}(\theta_{m,n})$ of Figure 1. The rectangles represent the 68 % HPD and the lines represent the 95 % HPD for the likelihood $\ell(y_{m,n}|\theta_{m,n})$ at each calibrated date.

2.3 Two prior models

Vertical mixing of earth is assumed to occur within layers, but not between layers. Thus objects found in lower layers are older than the objects found in superposed layers, whilst dates within a layer have no prior order constraint. See Section 4 for a more general setting. In our case study, $M = 6$ is the number of layers in the excavated series. For $m \in \{0, 1 \dots M\}$ let ψ_m denote the calendar date associated with the boundary between layers m and $m+1$. Layer $m = 1$ is the topmost and most recent layer, whilst layer $m = M$ is the deepest layer containing the oldest material. We have a total $K + M + 1$ unknown parameters: the $M + 1$ layer boundary dates $\psi_0 \dots \psi_M$, and the K unknown object dates, $\theta_{1,1} \dots \theta_{M,N_M}$. Let P and A, $P \leq A$ set lower and upper bounds on the otherwise unknown parameters. Possible parameter sets (ψ, θ) take some value in a parameter space $\Omega_{\Psi,\Theta}$. This space is simply the set of all states (ψ, θ) satisfying the stratigraphic constraints:

$$\Omega_{\Psi,\Theta} = \{(\psi, \theta); P \leq \psi_M \leq \theta_{M,\cdot} \leq \psi_{M-1} \leq \dots \psi_1 \leq \theta_{1,\cdot} \leq \psi_0 \leq A\}.$$

The posterior density $h(\psi, \theta|y, f)$ is defined in terms of an unnormalised prior density $f_{\Psi,\Theta}(\psi, \theta)$ say, and likelihood $L(y|\theta)$,

$$h(\psi, \theta|y, f) = L(y|\theta) \times f_{\Psi,\Theta}(\psi, \theta). \tag{4}$$

It is natural to model the object dates θ conditional on the layer boundary dates ψ, so we break our prior up in two pieces,

$$f_{\Psi,\Theta}(\psi, \theta) = f_{\Theta|\Psi}(\theta|\psi) f_{\Psi}(\psi).$$

In the absence of radiocarbon data, the parameters $\theta_{m,n}$ might take any value between ψ_m and ψ_{m-1} with equal probability, a state of knowledge represented by the choice

$$f_{\Theta|\Psi}(\theta|\psi) = \prod_{m=1}^{M} (\psi_{m-1} - \psi_m)^{-N_m}. \tag{5}$$

What density should we take for ψ ? A set of ψ values is in agreement with the stratigraphic data (so the ψ values are "legal") if $P \leq \psi_M \leq \psi_{M-1} \dots \leq \psi_0 \leq A$. It may seem natural to assert "any legal set of dates ψ is a priori equally likely", *ie*

$$f^{(1)}_{\Psi}(\psi) = 1$$

for all legal parameters sets ψ, and we will call this choice the *constant prior density*. In fact, the constant density for ψ weights the prior in favour of more widely spread sets of dates, and this can bias the whole analysis. The date span,

$$s(\psi) = \psi_0 - \psi_M, \tag{6}$$

which measures the number of years spanned by dated strata, exhibits the bias. Let $R = P - A$ denote the length of the interval between termini. Under the constant prior density, the marginal density of s is $(R - s)s^{M-1}$, so that a span of $2s$ is favoured over a span of s by a factor of about thirty (*ie*, 2^{M-1} when $s \ll R$) in our case study. This kind of weighting will not in general represent substantial prior knowledge. At the site in our case study, an alternative prior density,

$$f^{(2)}_{\Psi}(\psi) = s(\psi)^{1-M}/(R - s(\psi)),$$

under which spans of $2s$ and s are equally probable a priori, better represents prior knowledge. As an independent property, prior density $f^{(2)}_{\Psi}$ may be derived from a model of the

deposition and recovery processes as explained in Section 3. For $i = 1, 2$ let $f^{(i)}_{\Psi,\Theta}(\psi, \theta) = f_{\Theta|\Psi}(\theta|\psi) f^{(i)}_{\Psi}(\psi)$ denote priors derived under the two alternative ψ-models.

In setting up these models we have ignored the problem of inbuilt age in sample specimen. See Section 5 (final paragraph) for further discussion of this issue.

2.4 Comparison

The posterior densities determined by priors $f^{(1)}$ and $f^{(2)}$ and Equation (4) were summarised using samples generated using the rejection algorithms of Appendix A. All the software used to compute the graphs and Bayes factors presented here has been bundled together in a package, available at `car.ant.auckland.ac.nz`, and described in Jones and Nicholls 1999a. The rejection algorithm has the advantage, over an MCMC algorithm, that generated samples are IID. However, our rejection algorithm is suitable for small data sets only. Simulation from density $h(\psi, \theta|y, f^{(2)})$ with a large number of dates (say twenty) is a hard sampling problem even for MCMC, especially when states of small span are allowed by the likelihood. Simple MCMC algorithms (such as Gibbs samplers, simple random walk Metropolis Hastings and Metropolised rejection) suffer from convergence problems associated with the fact that the prior is unbounded. Details of a relatively complex Metropolis Hastings sampler adequate for all the large data sets we have encountered are given in Appendix C.

In Figure 3 the marginal posterior distributions of the span are presented. Notice that the posterior distribution of the span under prior $f^{(1)}$ is weighted towards larger values than is the case when prior $f^{(2)}$ is used. In Figure 4 the individual marginal posterior distributions of the calibrated dates are shown. Referring to the leftmost white rectangle, which is the posterior density of the date parameter $\theta_{1,1}$ of specimen (1,1), NZ 7758, computed using density $f^{(1)}$, we see that that parameter is concentrated at the upper end of its range. The posterior density of the date parameter $\theta_{6,1}$ of specimen (6,1), NZ 7771, computed using density $f^{(1)}$, is concentrated at the lower end of its range. Those parameters have bimodal posterior densities under $f^{(2)}$.

Which interpretation is favoured ? Bayesian model comparison of this kind is methodologically straightforward. See Carlin and Louis 1996 for background. We have two models, $\mathcal{M}_1$, corresponding to the prior choice $f^{(1)}$, and $\mathcal{M}_2$, for the $f^{(2)}$ prior. For model $\mathcal{M}_i$, $i = 1, 2$, let $P(\mathcal{M}_i|y)$ give the posterior probability, and let $P(\mathcal{M}_i)$ give the prior preference. The likelihood, $P(y|\mathcal{M}_i)$, for model $\mathcal{M}_i$ is

$$P(y|\mathcal{M}_i) \equiv \frac{\int_{\Omega_i} L(y|\theta) f^{(i)}_{\Psi,\Theta}(\psi, \theta) d\psi d\theta}{\int_{\Omega_i} f^{(i)}_{\Psi,\Theta}(\psi, \theta) d\psi d\theta} \tag{7}$$

with $\Omega_i = \Omega_{\Psi,\Theta}$ for $i = 1, 2$ in the comparison we are making. The *Bayes factor* $B(2\,v.1)$ for model $\mathcal{M}_2$ against model $\mathcal{M}_1$ is given by the ratio of the posterior probabilities for the two models,

$$B(2\,v.1) \equiv P(\mathcal{M}_2|y)/P(\mathcal{M}_1|y).$$

It follows, using Bayes rule, that under the non-informative choice $P(\mathcal{M}_1) = P(\mathcal{M}_2)$ of prior weighting, the ratio of the posterior probabilities for the two models equals the ratio of their likelihoods, $B(2\,v.1) = P(y|\mathcal{M}_2)/P(y|\mathcal{M}_1)$.

Computing the model likelihoods, using the estimators given in Meng and Wong 1996, (see

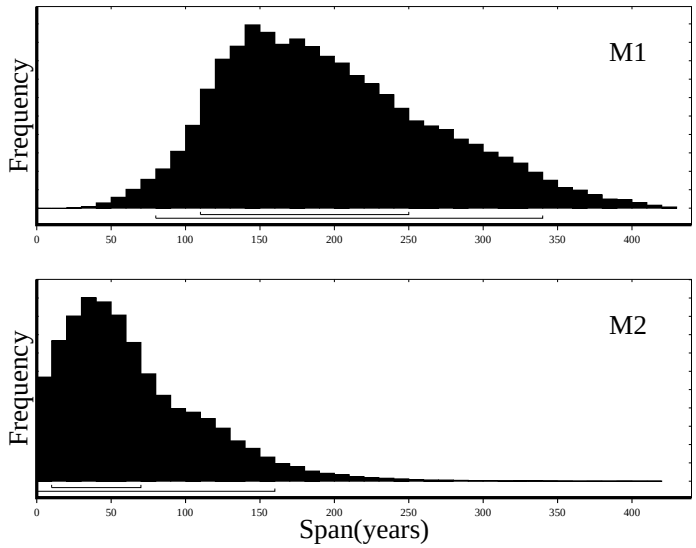

Figure 3: These histograms summarise the posterior distribution of the span, $s(\psi) = \psi_0 - \psi_M$, under prior models $f^{(1)}$ (graph $M1$, constant prior density for ψ) and $f^{(2)}$ (graph $M2$, uniform-span prior density for ψ). The posterior distribution of the span is sensitive to the choice of prior model.

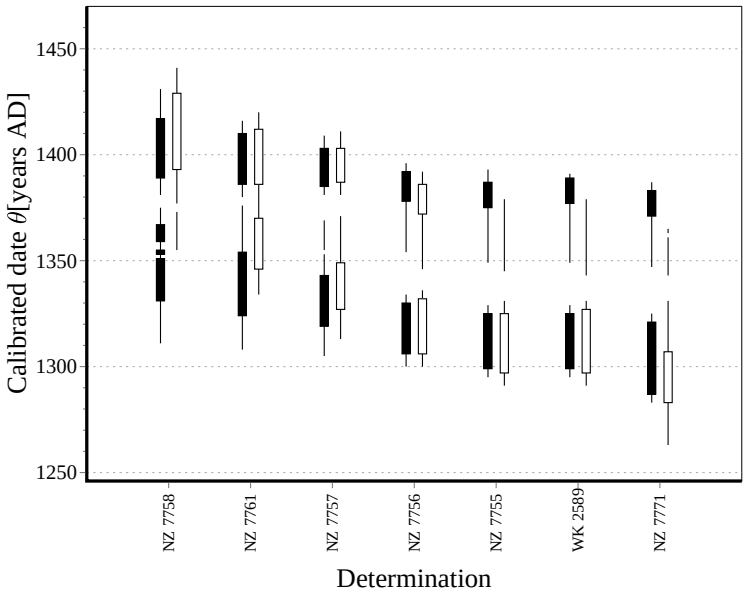

Figure 4: Marginal posterior distributions of $\theta_{m,n}$, the calibrated dates, are shown. The rectangles represent the 68 % HPD and the lines represent the 95 % HPD for the posterior count frequency. For each date the posterior under prior $f^{(1)}$ (white rectangles) and $f^{(2)}$ (black rectangles) are shown.

Appendix B for details) we find

$$P(y|\mathcal{M}_1) = 3.6(2) \times 10^{-19} \qquad P(y|\mathcal{M}_2) = 9.4(2) \times 10^{-18}$$

(standard errors in the place of the last quoted digit are given in parenthesis). The Bayes factor is $B(2\,v.1) = 26(2)$, so model $\mathcal{M}_2$ is about twenty six times more probable *a posteriori* than model $\mathcal{M}_1$. This confirms what we expect: the constant prior density of model $\mathcal{M}_1$ favours large span values, thereby pushing the calibrated dates θ towards values which are unrepresentative of the data. This leads to a small average likelihood for the constant prior density.

Prior $f^{(2)}$ gives a better summary than does $f^{(1)}$ of the prior knowledge available in this analysis. Also, because the span is a point of argument for this site, it is not appropriate to conduct an analysis with a prior weighted strongly toward large span values. Finally, the data supports prior density $f^{(2)}$ over $f^{(1)}$. We conclude that Figure 3-M1 is likely to overestimate the span of the events which generated the dates in our data set. Dates which have been calibrated using the constant prior density will be distorted in the same way.

3 A model of the deposition process

The distribution of the parameters of a prior should, where possible, be determined from a model of the physical processes which gave rise to the unknown true parameter values. In Christen *et al.* 1995, a deposition model for the specific case of peat accumulation with explicit depth data, plays the same role in the inference. Prior density $f^{(2)}_{\Psi,\Theta}(\psi,\theta)$ for specimen dates θ and layer boundary dates ψ may be derived from a model of the deposition and recovery processes which generate those specimen dates. The following should be regarded as an example of what can be done. Case-specific priors may be developed along the following lines for specific sites.

We model as follows. Parameters ψ_0 and ψ_M mark the start (ψ_M) and end (ψ_0) of human activity in the dated layers. Datable artifacts are generated according to a Poisson point process P_λ of piecewise constant rate $\lambda(t)$ for times t in the interval $[\psi_M, \psi_0]$. (A Poisson process is the neutral choice. It is equivalent to assuming that the process generates an event in any small interval of time with a probability which is proportional to the length of the interval and independent of the history of the process, to first order in the interval length.) Parameters $\psi_1 \ldots \psi_{M-1}$ mark change-points in $\lambda(t)$: these change points are themselves a realisation of a Poisson point process P_Λ of constant intensity in the interval $[\psi_M, \psi_0]$. Datable artifacts (generated by P_λ in our model) are randomly thinned by the natural processes of decay and by the specimen selection process of the archaeologist. These later processes are modelled as an independent geometric thinning of the datable artifacts generated by the deposition process. A Poisson process thinned in this way remains Poisson. The probability for a datable artifact to survive the thinning and become a dated specimen (*ie* yield a CRA) is assumed constant within layers, independent of the event that any other artifact survives, and independent of the age of the specimen given its layer (*ie* thinning can depend on age through layer).

Let $\tilde{f}_{\Psi,\Theta}(\psi,\theta)$ be the density of the (ψ,θ) values generated by these processes. This density breaks down into conditional components

$$\tilde{f}_{\Psi,\Theta}(\psi,\theta) = \tilde{f}_{\Theta|\Psi}(\theta|\psi)\tilde{f}_{\Psi_1\ldots\Psi_{M-1}|\Psi_0,\Psi_M}(\psi_1\ldots\psi_{M-1}|\psi_0,\psi_M)\tilde{f}_{\Psi_0,\Psi_M}(\psi_0,\psi_M).$$

We condition on knowledge of the number N_m of dated events in a layer. Since these events

are Poisson, their density is uniform in the interval in which they are generated,

$$\tilde{f}_{\Theta|\Psi}(\theta|\psi) = \prod_{m=1}^{M} (\psi_{m-1} - \psi_m)^{-N_m},$$

just as in Equation (5). We assume that all the change points generated by P_Λ are recorded by the archaeologist. We may then condition on knowledge of the number of events generated by P_Λ. As a consequence the $\psi_1 \dots \psi_{M-1}$ are again uniform,

$$\tilde{f}_{\Psi_1 \dots \Psi_{M-1}|\Psi_0,\Psi_M}(\psi_1 \dots \psi_{M-1}|\psi_0, \psi_M) \propto s(\psi)^{1-M}.$$

Our process model thus determines the prior density for all unknown parameters except ψ_0 and ψ_M.

We do not attempt to model the processes which determine the density $\tilde{f}_{\Psi_0,\Psi_M}(\psi_0, \psi_M)$. Instead we impose a weak bias towards shorter intervals with $\tilde{f}_{\Psi_0,\Psi_M}(\psi_0, \psi_M) = 1/(R-s(\psi))$ and thereby obtain the reference prior $\tilde{f}_{\Psi,\Theta} = f^{(2)}_{\Psi,\Theta}$. An alternative non-informative choice is $\tilde{f}_{\Psi_0,\Psi_M}(\psi_0, \psi_M) = 1$. This would lead to a marginal prior span density equal $(R-s)$. The functions we are considering here are relatively weak functions of s, whenever $s \ll R$. Among these mild reweightings we slightly favour $\tilde{f}_{\Psi_0,\Psi_M}(\psi_0, \psi_M) = 1/(R - s(\psi))$ because it gives uniform prior span, which is a convenient property (it is non-informative in respect of span, and may easily be reweighted). Could we not equally choose $\tilde{f}_{\Psi_0,\Psi_M}(\psi_0, \psi_M) = s(\psi)^{1-M}$, so that $\tilde{f}_{\Psi,\Theta}$ coincides with $f^{(1)}_{\Psi,\Theta}$? The factor $s(\psi)^{1-M}$ may vary a great deal over a small interval. It is highly informative with respect to (Ψ_0, Ψ_M) and, as we have seen, modifies the posterior distribution substantially. There is no information which supports that weighting for those variables. We conclude that the deposition-process model we have just described leads to a prior density like $f^{(2)}_{\Psi,\Theta}$ and unlike $f^{(1)}_{\Psi,\Theta}$.

In fact it is possible to find a process model which leads, in a natural way, to $f^{(1)}_{\Psi,\Theta}$. Suppose that $\psi_0, \psi_1 \dots \psi_M$ are all generated by a Poisson point process having constant intensity in $[P, A]$. This new process model is otherwise the same as the one above, so that the $\theta_{m,n}$ are again generated by a thinned Poisson process of intensity constant in $(\psi_{m-1}, \psi_m]$. The density for events (θ, ψ) generated by this new process is exactly $f^{(1)}_{\Psi,\Theta}$. The constant density prior $f^{(1)}_{\Psi,\Theta}$ is therefore equivalent to the assumption that all the boundary layer events in the model occur at a constant rate over the entire interval $[P, A]$. The constant prior tends to inflate the span of the reconstructed dates to fill the interval $[P, A]$, when $[P, A]$ is merely a conservative estimate, chosen so that P and A lie well outside the range of dates present in the analysis. A constant prior will be appropriate when P and A are given dates windowing a process which ran continuously before P and after A. This is all in contrast to the earlier model in which the events $\psi_1, \psi_2 \dots \psi_{m-1}$ occur at a constant rate in some sub interval $[\psi_0, \psi_M]$. In that model the date generation process is modelled as running for some interval of time starting after P and ending before A.

Consider the weaknesses in the deposition models we have outlined. The assumption that temporal deposition rates of datable artifacts are constant within layers is unreliable. The same problem may arise from sample selection bias. A layer typically contains many potentially datable artifacts. The archaeologist is often interested in the start and end dates for the layer and will, for that reason, select artifacts for dating that are likely to lie close to the extremes expected within the layer. This violates the assumptions we made about the thinning process. The selection may be biased in a number of other respects. Sensitivity to this assumption can be explored, by splitting such layers into two or three layers, and repeating the analysis. The constant prior density $f^{(1)}_{\Psi,\Theta}$ also has these faults.

4 Calibrated date and phase models

When the ψ parameters of the process model of Section 3 cannot be given the simple interpretation of that model, it may be best to drop the ψ variables from the model, but maintain the property of uniform prior span in the θ parameters. Models of calibrated dates (*ie* θ), with no layer parameters (no ψ-parameters), were common in the early literature (see for example Buck *et al.* 1991). They are particularly prone to show unrealistic spreading bias. The state space for this prior, Ω_Θ say, is simply the set of calibrated date sets θ which satisfy stratigraphic constraints. Thus in the case study of Section 2 we could remove the ψ variables and use

$$\Omega_\Theta = \{\theta; P \le \theta_{6,1} \le \theta_{5,2}, \theta_{5,1} \le \theta_{4,1} \le \theta_{3,1} \le \theta_{2,1} \le \theta_{1,1} \le A\}.$$

If $f_\Theta(\theta)$ is the prior density for a posterior density on Ω_Θ, take

$$s(\theta) = \max(\theta) - \min(\theta) \tag{8}$$

and use

$$f_\Theta(\theta) = s(\theta)^{2-K}/(R - s(\theta))$$

as reference prior. Under certain conditions this prior approximates the marginal of $f^{(2)}_{\Psi,\Theta}(\psi, \theta)$ after integration over ψ. We can show this explicitly when $M = 1$. Also, $f_\Theta(\theta)$ has uniform density on span, so it will often express the state of prior knowledge in radiocarbon date calibration. When this is not the case it makes a reasonable reference prior, with known properties, which can be modified so that it becomes representative of prior knowledge.

A setup like this, but using a prior $f_\Theta(\theta) = 1$, intended to be non-informative, is used in many places in the literature. This is reasonable when, as in Buck *et al.* 1991, the number of dates is very small. However, authors using $f_\Theta(\theta) = 1$ should note that they are asserting that a span of $2s$ is favoured over s by about 2^{K-2} as prior knowledge; the weighting depends now on K, the number of dates, which can be quite large, rather than on $M+1$, the number of layer variables, which is typically smaller. The resulting spreading bias can be extreme even in small problems.

The *multi-phase models* of Buck *et al.* 1992 include constant prior densities like $f^{(1)}$ as a special case. In multi-phase models, dated objects are associated with *phases*, which may overlap, rather than an abutting sequence of layers. For $m = 1, 2 \ldots M$, variables $\alpha_m, \beta_m \in [P, A], \alpha_m \le \beta_m$, with units calendar years AD, parameterise the start (α_m) and end (β_m) of the m'th phase. Phase parameters (α, β) constrain the calibrated dates in the same way that layer parameters do, with $\alpha_m \le \theta_{m,n} \le \beta_m$ for each $n = 1 \ldots N_m$. Constraints between phase parameters may be used to represent prior knowledge about the temporal order of the processes represented by the phases. Let $\Omega_{A,B} \subset [P, A]^{2M}$ be the space of a priori acceptable phase parameter sets in this model. If it is known that the start of phase m coincides with the end of phase $m + 1$, then we set $\alpha_m = \beta_{m+1}$. Constraints of this kind reduce the dimension of $\Omega_{A,B}$. Let d denote the number of phase parameters in the model (so $d = \dim(\Omega_{A,B})$). Note that, depending on the prior constraints defining $\Omega_{A,B}$, β_1 and α_M need not be the latest and earliest dates in a state. We must use

$$s(\alpha, \beta) = \max(\beta) - \min(\alpha)$$

to calculate the span.

What is a suitable prior here? We break the prior up as we did in Section 2.3, setting

$$f_{A,B,\Theta}(\alpha, \beta, \theta) = f_{\Theta|A,B}(\theta|\alpha, \beta) f_{A,B}(\alpha, \beta).$$

Prior ignorance of θ is represented by the choice

$$f_{\Theta|A,B}(\theta|\alpha,\beta) = \prod_{m=1}^{M} (\beta_m - \alpha_m)^{-N_m}.$$

Buck *et al.* 1996 then set $f_{A,B}(\alpha,\beta) = 1$ for each $(\alpha,\beta) \in \Omega_{A,B}$, and all later authors follow their choice. However, $f_\Delta(s) \propto (R-s)/s^{2-d}$ is then the marginal prior density of the span. If the number of degrees of freedom d is large then the constant phase prior again weights strongly in favour of large span. The uniform span density

$$f_{A,B}(\alpha,\beta) = s(\alpha,\beta)^{2-d}/(R - s(\alpha,\beta)) \tag{9}$$

makes a better reference prior. We have not motivated Equation (9) from a model of the processes generating its events. It does reduce to $f^{(2)}$ when the phases are in sequence and abutting ($\beta_{m-1} = \alpha_m$), and uniform span is desirable if prior knowledge is not informative in respect of this parameter. However, consider the following problem. Phase models form a large class. If we take the union of the events in two independent phase models the result is a phase model. We might expect the prior density to be the product of the densities for the two phase models. This is not what Equation (9) gives. An understanding of phase models as models of events generated by processes, generalising the model of Section 3 would be desirable.

5 Reservoir offsets and inbuilt-age

In this section we gather together three more or less unrelated points. Firstly, the setup used here (and in Buck *et al.* 1991) allows heterogeneous date types to be calibrated together. Thus if $y_{m,n}$ is an obsidian hydration date, we simply replace the likelihood factor $\ell(y_{m,n}|\theta_{m,n})$ by the appropriate density for that observation process.

Secondly, the error analysis of so called "reservoir offsets" is universally poorly handled. The atmosphere is a kind of carbon reservoir. Since its radiocarbon activity varies spatially, calibration data are strictly applicable to the immediate carbon reservoir from which the calibrating material was sourced. However, it is widely assumed that any spatial variation in the calibration data can be modelled as a reservoir offset Δ which does not vary in time. The atmospheric activity in New Zealand is considered to differ from that for which the standard calibration data has been developed (*e.g.* McCormac *et al.* 1998). Thus if $\mu_{\rm NZ}$ is the New Zealand calibration function, then $\mu_{\rm NZ}(\theta) = \mu(\theta) - \Delta_{\rm NZ}$, and McCormac *et al.* 1998 report $\Delta_{\rm NZ} = -27$ with a $\sigma^{(\Delta)} = 5$ standard error. The offset model we have just described is one in which all CRA's y in the observation model of Equation (1) suffer an additional *common* random offset δ, so that

$$y_{m,n} = \mu(\Theta_{m,n}) - \delta + \epsilon^{(Y)}_{m,n} + \epsilon^{(\mu)}_{m,n}|\Theta_{m,n}$$

with $\delta \sim \mathsf{N}(\Delta, {\sigma^{(\Delta)}}^2)$. The unknown offset is a nuisance parameter in the likelihood,

$$L(y,\Delta|\theta,\delta) = \frac{e^{-(\delta-\Delta)^2/2{\sigma^{(\Delta)}}^2}}{\sqrt{2\pi{\sigma^{(\Delta)}}^2}} \prod_{m=1}^{M}\prod_{n=1}^{N_m} \ell(y_{m,n}+\delta|\theta_{m,n}).$$

Uncertainty in offset measurements has until now been treated by setting

$$\sigma(\Theta_{m,n})^2 \stackrel{?}{=} \sigma^{(\mu)}(\Theta_{m,n})^2 + {\sigma^{(Y)}_{m,n}}^2 + {\sigma^{(\Delta)}}^2$$

and replacing each CRA $y_{m,n}$ by $y_{m,n}+\Delta$ in the subsequent analysis. This gives *independent* random offsets,

$$y_{m,n} \stackrel{?}{=} \mu(\Theta_{m,n}) - \delta_{m,n} + \epsilon^{(Y)}_{m,n} + \epsilon^{(\mu)}_{m,n} | \Theta_{m,n}$$

with $\delta_{m,n} \sim \mathsf{N}(\Delta, {\sigma^{(\Delta)}}^2)$ realised independently for each CRA $y_{m,n}$, $m = 1, 2 \ldots M$ and each $n = 1, 2 \ldots N_m$. We do not believe this is intended. The practical impact of this error is generally slight since $\sigma^{(\Delta)}$ is usually small compared to $\sigma^{(Y)}_{m,n}$. The sensitivity of southern hemisphere dating results to this observation model misspecification is investigated further in Jones and Nicholls 1999b. In the case study of Section 2 no offset is used.

Finally, we come to the problem of "in-built age". Radiocarbon measurement $y_{m,n}$ dates the death of the material comprising dated specimen (m, n). It is deposited in its archeaological context some time after dying. It is possible that the context date, $\phi_{m,n}$ say, significantly post-dates the event dated by radiocarbon. Certainly we can usually assume that $\phi_{m,n} \geq \theta_{m,n}$. Heart wood from large trees may have millenia of in-built age. It follows that radiocarbon event dates need not observe the temporal order suggested by stratigraphy. In practice it is usual to assume that the formation of the archaeological context and the dated event are isochronous. The data in Table 1 come from short lived species of shrub, so it is expected that dates for radiocarbon event (demise of shrub) and context (buried hearth) are not separated by any significant interval of time. However, there will be cases in which the event and context dates are not isochronous. The likelihood is unchanged by these considerations. If the measured event date $\theta_{m,n}$ is independent of all layer boundary dates ψ once we condition on the context date $\phi_{m,n}$, then we can break up the new prior density, $f_{\Psi,\Phi,\Theta}(\psi, \phi, \theta)$ say, as

$$f_{\Psi,\Phi,\Theta}(\psi, \phi, \theta) = f_{\Theta|\Phi}(\theta|\phi) f_{\Phi|\Psi}(\phi|\psi) f_{\Psi}(\psi).$$

Here the functions $f_{\Phi|\Psi}(\phi|\psi)$ and $f_{\Psi}(\psi)$ are identical, respectively, to $f_{\Theta|\Psi}(\theta|\psi)$ and $f_{\Psi}(\psi)$ of Section 2.3 (so ϕ here takes the place of θ in that prior). Assuming conditional independence of event dates given context dates, we can model $f_{\Theta_{m,n},\Phi_{m,n}}$ as a function of $\phi_{m,n} - \theta_{m,n}$ if the dated material type is known. For example, for CRA's from wood of known species, explicit modelling of tree growth rate and shape supports an exponential density $f_{\Theta_{m,n}|\Psi_{m,n}}(\theta_{m,n}|\psi_{m,n}) \propto \exp(-(\phi_{m,n} - \theta_{m,n})/\gamma)$, whose mean γ is related to the tree lifespan and growth habit.

6 Discussion and Conclusions

Before concluding, we return to the Stein effect we mentioned in the introduction. A Bayesian derivation of the James-Stein shrinkage estimator ($\hat{\theta}^{\mathrm{JS}}_{m,n}$ say) for the unknown mean parameters $\Theta_{m,n}$ of observations of independent normal variates $y_{m,n} \sim N(\Theta_{m,n}, c^2)$, with c a known constant, is given, for example, in Carlin and Louis 1996 pages 84-88. The unknown means $\Theta_{m,n}$ are assumed to be independently realised from a normal prior distribution with unknown mean and variance. Estimates $\hat{\theta}^{\mathrm{JS}}_{m,n}$ are squeezed by this overall prior distribution toward the mean of the observations. Our Θ-parameters are thinned realisations of nested Poisson processes. However, the basic setup is the same and shrinkage occurs for the same reasons. Our analysis is complicated by the presence of a non-linear function μ in the observation process, and by stratigraphic constraints. As a consequence we are not able to get a general Stein-like point estimator for Θ in closed form, though a sample based estimate is easily obtained for any particular dataset.

Under what circumstances is it appropriate to use a uniform-span prior? A small number of outlier values will in certain cases be shrunk excessively towards a larger central pool of

dates. In a similar fashion, James-Stein estimators have a lower total mean squared error loss than the maximum likelihood estimator even though the mean squared loss on a single parameter can be much larger. The rule for James-Stein estimators applies here: squeeze a set of dates when it is appropriate to spread the error loss through the whole set. We refer readers to the literature on James-Stein estimators.

Under what circumstances does a uniform-span prior make a difference ? To the best of our knowledge, replacing $f^{(1)}$ by $f^{(2)}$ makes no significant difference for data sets of the kind considered in the Bayesian radiocarbon calibration literature up to the present. The full data set discussed in Anderson *et al.* 1996, of which our example is a small part, is highly sensitive. However, a Bayesian analysis of that data has not been published. The radiocarbon likelihood Equation (3) has a span dependence which is approximately Gaussian (at least over large date ranges) whilst the weighting of the prior $f^{(1)}_{\Psi,\Theta}$ (relative to $f^{(2)}_{\Psi,\Theta}$) towards greater spans is a power. It follows that spreading due to $f^{(1)}$ will be visible in so far as it is admitted by the likelihood. Such spreading is noticeable when the standard errors $\sigma(\theta)$ approach in magnitude the span of values seen in y, the CRA data.

Can one interpret ψ_0 and ψ_M literally as dates marking the start and end of activity in the series under study? This is the case when the model of Section 3 gives a good description of the processes which generated the dated specimens. One is then able to make simple summarising statements about the processes under study. In particular, the *absence* of datable specimen becomes informative. When the deposition model is quite unrealistic the ψ variables should either be dropped (density f_Θ and parameter space Ω_Θ of Section 4 replace $f_{\Psi,\Theta}$ and parameter space $\Omega_{\Psi,\Theta}$ of Section 2.3) or they should be interpreted as irrelevant auxiliary variables used to form a non-informative prior for θ. Whether ψ variables are present or not we should, in this case, use the revised definition of span $s(\theta)$ given in Equation (8). This quantity estimates a lower bound on the duration of the activity which generated the dated specimens, rather than estimating the duration itself.

The posterior distribution of the span is sensitive also to errors in the measurement of the standard errors $\sigma^{(Y)}_{m,n}$ which accompany a CRA. In this paper, as elsewhere, $\sigma^{(Y)}_{m,n}$ are assumed to be known parameters of the observation process, when in fact they are measured. If the $\sigma^{(Y)}_{m,n}$ are overestimated then the posterior distribution of the span is shifted to unrepresentatively small values. If the $\sigma^{(Y)}_{m,n}$ are underestimated, the reverse is true.

The usual caveats of Bayesian modelling apply: prior densities of meaningful quantities should be sumarised via histograms or if possible analytically to check that these marginal densities represent prior information reasonably well. Buck *et al.* 1991 illustrate this procedure. When there is doubt about which model to use, Bayes factors may be helpful. Bayes factors have reputation for being easy to interpret but hard to estimate. However we have found estimation with the Meng-Wong estimator (see Appendix B and Meng and Wong 1996) quite straightforward in all radiocarbon calibration problems we have encountered.

References

Anderson, A. J., Smith, I. W. G., and Higham, T. F. G. 1996. Shag river mouth: the archaeology of an early Southern Maori village. *Pages 61–69 of:* Anderson, A. J., Allingham, B., and Smith, I. W. G. (eds), *Shag river mouth*, vol. 27. ANU, Canberra, Australia: ANH Publications, RSPAS.

Buck, C.E., Kenworthy, J.B., C.D.Litton, and Smith, A.F.M. 1991. Combining archaeolog-

ical and radiocarbon information: a Bayesian approach to calibration. *Antiquity*, **65**, 808–821.

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.

Buck, C.E., Cavanagh, W.G., and C.D.Litton. 1996. *Bayesian approach to interpreting archaeological data*. Chichester: John Wiley.

Buck, C.E., Christen, J.A., and James, G.N. 1999. Bcal: an online Bayesian radiocarbon calibration tool. *Internet archaeology*, **7**. `http://intarch.ac.uk/`.

Carlin, B.P., and Louis, T.A. 1996. *Bayes and empirical Bayes methods for data analysis*. London: Chapman and Hall.

Christen, J.A. 1994. Summarizing a set of radiocarbon determinations: a robust approach. *Appl. Statist.*, **43**, 489–503.

Christen, J.A., Clymo, R.S., and Litton, C.D. 1995. A Bayesian approach to the use of C14 dates in the estimation of the age of peat. *Radiocarbon*, **37**, 425–430.

Jones, M., and Nicholls, G.K. 1999a. *New date calibration software*. in preparation.

Jones, M., and Nicholls, G.K. 1999b. *On quantifying uncertainty due to hemisphere offset errors*. Tech. rept. Center for Archaeological Research, The University of Auckland.

Litton, C., and Buck, C. 1996. *Markov chain Monte Carlo in practice*. Chapman and Hall. Chap. 25, pages 466–486.

McCormac, F.G., Hogg, A.C., Higham, T.F.G., Baillie, M.G.L., Palmer, J.G., Xiong, L., Pilcher, J.R., Brown, D., and Hoper, S.T. 1998. Variations of radiocarbon in tree-rings: Southern hemisphere offset preliminary results. *Radiocarbon*, **40**(3), 1153.

Meng, X-L., and Wong, W.H. 1996. Simulating ratios of normalising constants via a simple identity: a theoretical exploration. *Statistica Sinica*, **6**, 831–860.

Pauler, D. K., Wakefield, J. C., and Kass, R. E. 1997. *Bayes factors for variance component models*. Tech. rept. Department of Statistics, Carnegie Mellon University, `http://lib.stat.cmu.edu/cmu-stats/tr/tr669/tr669.html`.

Ramsey, C.B. 1995. Radiocarbon calibration and analysis of stratigraphy: The OxCal program. *Radiocarbon*, **37**, 425–430.

Stein, C. 1956. Inadmissability of the usual estimator for the mean of a multivariate normal distribution. *Pages 197–206 of: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* University of California Press, for Berkeley.

Stuiver, M., Reimer, P.J., E. Bard, J.W. Beck, Burr, G.S., Hughen, K.A., Kromer, B., McCormac, F.G., Plicht, J.v.d., and Spurk, M. 1998. Intcal98 radiocarbon age calibration, 24,000-0 cal BP. *Radiocarbon*, **40**, 1041–1083.

Tierney, L. 1996. *Markov chain Monte Carlo in practice*. Chapman and Hall. Chap. 4, pages 59–74.

Simulation for $f_\Psi = f_\Psi^{(2)}$

$s \sim \mathsf{Uniform}(0, R)$

$\psi_M \sim \mathsf{Uniform}(P, A - s)$

$\psi_0 = \psi_M + s$

for each $m \in \{1, 2 \dots M-1\}$

$\quad x_m \sim \mathsf{Uniform}(\psi_M, \psi_0)$

$[\psi_1 \dots \psi_{M-1}] \leftarrow \mathtt{sort}(x_1 \dots x_{M-1})$

return ψ

Simulation for $f_\Psi = f_\Psi^{(1)}$

for each $m \in \{0, 1, \dots M\}$

$\quad x_m \sim \mathsf{Uniform}(P, A)$

$[\psi_0 \dots \psi_M] \leftarrow \mathtt{sort}(x_0 \dots x_M)$

return ψ

Rejection algorithm returning $(\psi, \theta) \sim h(\psi, \theta | y, f)$

repeat

$\quad \psi \sim f_\Psi(\psi)$

$\quad u \sim \mathsf{Uniform}(0, 1)$

until $u \leq h_\Psi(\psi | y, f) / c f_\Psi(\psi)$

for each $(m, n) \in \{(1, 1) \dots (M, N_M)\}$

$\quad \theta_{m,n} \sim h_{\Theta|\Psi}(\theta | \psi, y)$

return (ψ, θ)

Figure 5: Rejection algorithm returning samples $(\psi, \theta) \sim h(\psi, \theta | y, f)$. The marginal prior density $f_\Psi(\psi)$ is simulated using to the appropriate function above, depending on whether prior model $f = f^{(1)}$ or $f = f^{(2)}$ is used. The function $\psi = \mathtt{sort}(x)$ sorts the elements of the vector x into descending order. The quantity c is defined in Equation (10). The simulations $\theta_{m,n} \sim h_{\Theta_{m,n}|\Psi,Y}(\theta_{m,n} | \psi, y)$, which might themselves be carried out by rejection (or numerical inversion) are not detailed.

Appendix A: Rejection sampling algorithm

In this section we describe an algorithm returning samples distributed according to the posterior density. This algorithm is adequate for small data sets only.

Let $\mathbb{I}_{x\in S}$ equal the indicator function for the event that x is a member of the set S. The density $h(\psi,\theta|y,f)$ breaks down as

$$h(\psi,\theta|y,f) = h_\Psi(\psi|y,f)\prod_{m=1}^{M}\prod_{n=1}^{N_m} h_{\Theta_{m,n}|\Psi,Y}(\theta_{m,n}|\psi,y),$$

where $h_\Psi(\psi|y,f)$ is the marginal posterior density for the ψ dates

$$h_\Psi(\psi|y,f) = f_\Psi(\psi)\prod_{m=1}^{M}\prod_{n=1}^{N_m}\frac{\int_{\psi_m}^{\psi_{m-1}}\ell(y_{m,n}|\theta_{m,n})d\theta_{m,n}}{\psi_{m-1}-\psi_m},$$

and

$$h_{\Theta_{m,n}|\Psi,Y}(\theta_{m,n}|\psi,y) \propto \ell(y_{m,n}|\theta_{m,n}) \times \mathbb{I}_{\theta_{m,n}\in[\psi_m,\psi_{m-1}]}$$

We simulate $\psi \sim h_\Psi(\psi|y,f)$ *via* rejection, using $f_\Psi(\psi)$ as the envelope (covering) function. We then simulate $\theta_{m,n} \sim h_{\Theta_{m,n}|\Psi,Y}(\theta_{m,n}|\psi,y)$, using their conditional independence. We bound the ratio $h_\Psi(\psi|y,f)/f_\Psi(\psi)$ by noting that if

$$c = \prod_{m=1}^{M}\prod_{n=1}^{N_m}\max_{z\in[P,A]}\ell(y_{m,n}|z) \tag{10}$$

then

$$h_\Psi(\psi|y,f) \le c f_\Psi(\psi).$$

Algorithmic details are given in Figure 6

Appendix B: estimation of Bayes factors

How are the Bayes factors given in Section 2.4 actually estimated ? Let

$$\mathcal{Z}_h = \int_\Omega L(y|\theta) f_{\Psi,\Theta}(\psi,\theta)d\psi d\theta$$

and let

$$\mathcal{Z}_f = \int_\Omega f_{\Psi,\Theta}(\psi,\theta)d\psi d\theta$$

We need to estimate the ratio of posterior to prior normalising constants, $P(y|\mathcal{M}) = \mathcal{Z}^h/\mathcal{Z}^f$, for each model. After some searching we found a stable and efficient estimator: the simplest "importance sampling" estimator reported in Meng and Wong 1996. When samples are iid, the standard error can be computed directly, without resampling.

Let $S_f = \{\psi^{(f,i)},\theta^{(f,i)}\}_{i=1}^{J_f}$ be a set of iid samples distributed according to the prior density, $f_{\Psi,\Theta}$, and let $S_h = \{\psi^{(h,i)},\theta^{(h,i)}\}_{i=1}^{J_h}$ be a set of iid samples distributed according to $h_{\Psi,\Theta}$, the posterior density. Let $r = \mathcal{Z}^h/\mathcal{Z}^f$ and let $\hat{r}_{MW}(S)$ denote the Meng-Wong estimator

for r computed from the sample set $S = \{S_f, S_h\}$. An iterative sequence of estimates, $\hat{r}_t$ with $t = 1, 2 \ldots$, is given below. Let $J = J_f + J_h$, let

$$E_f(\hat{r}_t|S) \equiv \frac{J}{J_f} \sum_{i=1}^{J_f} \frac{L(y|\theta^{(f,i)})}{J_h L(y|\theta^{(f,i)}) + J_f \hat{r}_t}$$

and let

$$E_h(\hat{r}_t|S) \equiv \frac{J}{J_h} \sum_{i=1}^{J_h} \frac{1}{J_h L(y|\theta^{(h,i)}) + J_f \hat{r}_t}.$$

The iteration determining $\hat{r}_{MW}$ is

$$\hat{r}_{t+1} = E_f(\hat{r}_t|S)/E_h(\hat{r}_t|S).$$

The sequence $\hat{r}_1, \hat{r}_2 \ldots$ converges to the unique number $\hat{r}_{MW}$ from any starting value $\hat{r}_1$. We initialise the sequence with the Gelman estimator $\hat{r}_1 = \hat{r}_{GM}$ for r,

$$\hat{r}_{GM} = \frac{\frac{1}{J_f} \sum_{i=1}^{J_f} h(\psi^{(f,i)}, \theta^{(f,i)}|y)}{\frac{1}{J_h} \sum_{i=1}^{J_h} f(\psi^{(h,i)}, \theta^{(h,i)})}.$$

Let $\epsilon^2 = \mathsf{E}_S\{(\hat{r}_{MW}(S) - P(y))^2\}$ and let $\hat{\epsilon}$ estimate ϵ. Meng and Wong give an error estimate,

$$\frac{\hat{\epsilon}^2}{\hat{r}_{MW}^2} = \frac{J_f J_h}{J} \times \frac{1}{E_f(\hat{r}_{MW})} - \frac{J}{J_f J_h},$$

to first order in J, the total number of samples.

When samples are obtained by rejection, with the proposal function proportional to the prior, (or a marginal density of the prior as in Appendix A) one has the neat result that, if s samples are realised from a total of b rejection sampling draws, then $\hat{r}_R = s/b$ is an estimate for r with variance $r(r-1)/b$. Starting with this result, Pauler *et al.* 1997 give an estimator for r having smaller variance. We used this simple "rejection count" estimator to check estimates $\hat{R}_{MW}$ obtained using our implementation of the Meng-Wong estimator.

Appendix C: Metropolis-Hastings sampling algorithm

We implemented a Metropolis Hastings MCMC algorithm to generate samples ψ, θ distributed with density $h(\psi, \theta|y, f)$ for $f = f^{(1)}$ and $f = f^{(2)}$.

We define an algorithm generating a realisation $\{\psi^{(j)}, \theta^{(j)}\}_{j=0}^{J}$ of a Markov chain of random variables $\{\Psi^{(j)}, \Theta^{(j)}\}_{j=0}^{J}$, with equilibrium density $h_{\Psi,\Theta}$. In order to get an efficient sampler, it is effective to use several transition rules, labelled $v = 1, 2 \ldots$, each with its own transition kernel. Each rule is made up of a generation step, in which a candidate state (ψ', θ') is generated according to a density q_v for rule v, and an acceptance step, in which the candidate state is accepted with probability $\alpha_v(\psi', \theta'|\psi, \theta)$, or rejected. The algorithm is as follows.

Let $(\Psi^{(j)}, \Theta^{(j)}) = (\psi, \theta)$. $(\Psi^{(j+1)}, \Theta^{(j+1)})$ is determined in the following way.

1. Pick a new state $(\psi', \theta') \sim q_v(\psi'\, \theta'|\psi, \theta)$.

2. Accept (ψ', θ') (*ie*, set $(\Psi^{(j+1)}, \Theta^{(j+1)}) = (\psi', \theta')$) with probability

$$\alpha_v(\psi', \theta' | \psi, \theta, f) \equiv \min \left\{ 1, \frac{h(\psi'\theta'|y, f) q_v(\psi\,\theta|\psi', \theta')}{h(\psi\,\theta|y, f) q_v(d\psi' d\theta'|\psi, \theta)} \right\}.$$

If (ψ', θ') is not accepted, set $(\Psi^{(j+1)}, \Theta^{(j+1)}) = (\psi, \theta)$ (*ie*, no change).

Sufficient conditions for (Ψ, Θ) to have equilibrium distribution h are given, for example, in Tierney 1996. We have four update types, with distinct generation distributions q_v, so the above algorithm is repeated for each update type, $v = 1, 2, 3, 4$ in sequence, and the sequence repeated until enough samples have been generated.

The first three update types are quite standard. We give the candidate generartion procedure only. The fourth type is particularly effective for MCMC mixing. It is hard to compute the acceptance probability so we give more detail.

- Update a single θ variable, $v = 1$, $(\psi, \theta \to \psi, \theta')$.
  A parameter $\theta_{m,n}$ is chosen uniformly at random (UAR) from the set $\{\theta_{1,1} \dots \theta_{N,N_M}\}$. A new value for $\theta_{m,n}$ ($\theta'_{m,n}$ say) is selected UAR on the interval $[\psi_m, \psi_{m-1}]$.

- Update a single ψ variable, $v = 2$, $(\psi, \theta \to \psi', \theta)$.
  An index m is chosen UAR in $0, 1 \dots M$. A new value of ψ'_m for ψ_m is selected UAR on the interval $[\max(\theta_{m+1,\cdot}), \min(\theta_{m,\cdot})]$ (the interval is $[L, \min(\theta_{1,\cdot})]$ if $m = 0$ and $[\max(\theta_{M-1,\cdot}), U]$ if $m = M$).

- Shift all dates, $v = 3$, $(\psi, \theta \to \psi', \theta')$.
  In this update S is a positive constant with units *years* chosen to give a reasonable acceptance rate for the update. A scalar shift s is chosen uniformly at random in $[-S, S]$ and added to each variable.

- Expand the dates about their mean, $v = 4$, $(\psi, \theta \to \psi', \theta')$.
  A scalar multiplying factor ρ is chosen uniformly at random in $[2/3, 3/2]$. Let $\mathrm{av}(\psi, \theta)$ denote the arithmetic mean of the sequence ψ, θ of dates. We set $\theta'_{m,n} = \rho\theta_{m,n} - (\rho - 1)\mathrm{av}(\psi, \theta)$ and $\psi'_m = \rho\psi_m - (\rho - 1)\mathrm{av}(\psi, \theta)$, for each $m = 0 \dots M$ and $n = 1 \dots N$. This operation scales all the dates by ρ whilst keeping $\mathrm{av}(\psi, \theta) = \mathrm{av}(\psi', \theta')$. This time the correct acceptance rates are difficult to calculate. We find

$$\alpha_4(\psi', \theta' | \psi, \theta, f^{(1)}_{\Psi,\Theta}) = \min \left\{ 1, \frac{L(y|\theta')}{L(y|\theta)} \times \rho^{M-2} \right\}$$

and

$$\alpha_4(\psi', \theta' | \psi, \theta, f^{(2)}_{\Psi,\Theta}) = \min \left\{ 1, \frac{L(y|\theta')}{L(y|\theta)} \times \frac{R - (\psi_0 - \psi_M)}{R - \rho(\psi_0 - \psi_M)} \times \frac{1}{\rho} \right\}$$

with

$$\frac{L(y|\theta')}{L(y|\theta)} = \frac{e^{-|y-\mu(\theta')|^2/2\sigma(\theta')^2}}{e^{-|y-\mu(\theta)|^2/2\sigma(\theta)^2}} \times \prod_{m=0}^{M} \prod_{n=1}^{N} \frac{\sigma_{m,n}(\theta_{m,n})}{\sigma_{m,n}(\theta'_{m,n})}$$

The posterior $h_{\Psi,\Theta}(\psi\,\theta|y, f^{(1)}_{\Psi,\Theta})$ may in general be simulated adequately with moves $v = 1$ and $v = 2$ alone. However, in order to simulate reliably the posterior with prior density $f(2)_{\Psi,\Theta}$, we require in addition move $v = 4$. Move $v = 3$ improves the efficiency of the sampling process.