NEW METHODS AND CRITICAL ASPECTS IN BAYESIAN MATHEMATICS FOR ^{14}C CALIBRATION

Peter Steier[1] • Werner Rom[2] • Stephan Puchegger[1]

ABSTRACT. The probabilistic radiocarbon calibration approach, which largely has replaced the intercept method in ^{14}C dating, is based on the so-called Bayes' theorem (Bayes 1763). Besides single-sample calibration, Bayesian mathematics also supplies tools for combining ^{14}C results of many samples with independent archaeological information such as typology or stratigraphy (Buck et al. 1996). However, specific assumptions in the "prior probabilities", used to transform the archaeological information into mathematical probability distributions, may bias the results (Steier and Rom 2000). A general technique for guarding against such a bias is "sensitivity analysis", in which a range of possible prior probabilities is tested. Only results that prove robust in this analysis should be used. We demonstrate the impact of this method for an assumed, yet realistic case of stratigraphically ordered samples from the Hallstatt period, i.e. the Early Iron Age in Central Europe.

INTRODUCTION

From a radiocarbon measurement one can easily get a mathematical probability distribution, i.e. a completely quantitative result. Regarding the dating of an archaeological sample, this distribution for every calendar year contains the probability that the sample originated in that individual year. In many cases the ^{14}C age is not the only available information on archaeological samples. Additional information may originate from typology, stratigraphy, or dendrochronology. Typology and stratigraphy supply information in a qualitative form, i.e. they allow grouping of samples or determining their relative chronological order. By accounting for this additional information one may expect to achieve a better-defined date than by using only the information from the ^{14}C measurement.

In the conventional way, these two types of information are combined as follows: the probability distributions from the ^{14}C measurement are condensed into 95% highest probability density (HPD) intervals, i.e. intervals on the calendar axis that contain the most probable 95% of the respective distributions; next, this condensed information is combined with the thereof independent additional knowledge by archaeological reasoning. However, several insufficiencies of this approach are obvious: on one hand, information contained in the details of the probability distribution is thrown away. On the other hand, the region containing the residual 5% that lie outside the 95%-HPD interval will be simply ignored, although this region may become relevant when the ^{14}C age information is combined with the additional archaeological information.

One may overcome both these insufficiencies by using Bayesian statistics (Bayes 1763; Jeffreys 1961), a method that has been successfully applied in many scientific fields (Malakoff 1999). In general, Bayesian statistics are a consistent mathematical framework to update *prior* information or knowledge with new data from a measurement to arrive at a more accurate *posterior* knowledge. This, yet, requires that all information is available as probability distributions. So the Bayesian approach in ^{14}C dating is different from conventional archaeological reasoning insofar as the qualitative archaeological information must be transformed into mathematical probability distributions for the calendar ages of the samples. For a set of samples *A*, *B*,... with calendar ages t_A, t_B,..., we identify the additional archaeological information with the *prior probability*, and the calibrated ^{14}C probability distributions $P_A^{cal}(t_A)$, $P_B^{cal}(t_B)$,... with the *new data*. By applying Bayes' theorem we

[1]Vienna Environmental Research Accelerator, Institut für Isotopenforschung und Kernphysik, Universität Wien, Währinger Straße 17, A-1090 Wien, Austria. Email: peter.steier@univie.ac.at.
[2]AMS ^{14}C Dating Laboratory, Institut for Fysik og Astronomi Aarhus Universitet, DK-8000 Ärhus C, Denmark

RADIOCARBON, Vol 43, Nr 2A, 2001, p 373–380
Proceedings of the 17th International ^{14}C Conference, edited by I Carmi and E Boaretto

obtain "updated" probability distributions, the *posterior probabilities*, and especially the *marginal posterior probability distributions*, $P_A^{posterior}(t_A)$, $P_B^{posterior}(t_B)$, ... for the samples. In the simplest case of only two samples *A* and *B*, by using Bayes' theorem these marginal posterior probability distributions can be written as:

$$P_A^{posterior}(t_A) = \frac{1}{U_A}\int P_A^{cal}(t_A)\cdot P_B^{cal}(t_B)\cdot P^{prior}(t_A, t_B)dt_B$$

$$P_B^{posterior}(t_B) = \frac{1}{U_B}\int P_A^{cal}(t_A)\cdot P_B^{cal}(t_B)\cdot P^{prior}(t_A, t_B)dt_A\,. \qquad (1)$$

U_A and U_B are constants needed to normalize the $P^{posterior}$ to unity.

TRANSFORMING ARCHAEOLOGICAL INFORMATION INTO MATHEMATICAL PROBABILITY DISTRIBUTIONS

The critical step in the Bayesian approach is the conversion of the additional archaeological information into mathematical probability distributions, which then serve as the prior. A complete prior for *N* samples requires that a probability $P_{prior}(t_1,...,t_N)$ is assigned to every combination of calendar dates $t_1,..., t_N$ of the *N* samples.

We ignore the advantageous cases where there exists a straightforward method to construct the prior (e.g. tree-ring wiggle matching [Bronk Ramsey 2001] or the individual calibration of ^{14}C samples, but see Section "A subtlety in single-sample calibration" below) and focus on the case of stratigraphy. We study the simplest case of two samples *A* and *B*, where stratigraphy infers that the (true) calendar age t_A of sample *A* is higher than the (true) calendar age t_B of sample *B*. We can express this by the following (generic) probability distribution:

$$P^{prior}(t_A, t_B) = \begin{cases} 0 & if\ t_A \le t_B\ (\text{"}forbidden\text{"}\ case) \\ ? & if\ t_A > t_B\ (\text{"}allowed\text{"}\ case) \end{cases} . \qquad (2)$$

$P^{prior}(t_A, t_B)$ is well defined for "forbidden" age combinations which are in contradiction to the stratigraphic evidence: The probability for these age combinations is 0. Concerning the "allowed" region, an archaeologist who is also familiar with Bayesian mathematics may supply us with an educated guess by means of "subjective" prior elicitation methods (Berger 1980). However, the use of such "subjective" priors is at dispute in the mathematical literature (Malakoff 1999). A constant probability in any "allowed" region (as used by Bayes and later by Laplace) is the simplest assumption, but in many applications these so-called "uniform" priors perform badly since they bias the posterior results (Jeffreys 1961; Yang and Berger 1997). Modern Bayesian statistics has therefore developed a variety of methods for deriving so called "noninformative" priors, which are constructed such as to have minimal biasing on the final result (Berger 1980).

Concerning the ^{14}C calibration of two stratigraphically ordered samples, the uniform prior is the only prior used and studied so far. In Steier and Rom (2000) we demonstrated by means of computer-simulated experiments that this prior tends to increase the age differences of ordered samples that are not clearly separated by the ^{14}C measurement. One may try now to find a better noninformative prior, but we adopt a different method to deal with this problem.

https://doi.org/10.1017/S0033822200038236 Published online by Cambridge University Press

Robust Bayesian Analysis

The basic idea of "robust Bayesian analysis" (see Berger 1994 for an overview) is to test a variety of priors and by this to check their influence on the answer for a given problem. This method is especially suited to handle cases where the result depends significantly on the selected prior, i.e. where the answer is highly uncertain and conclusions therefore cannot be accurate. Only from posterior results which are robust with respect to the varied priors accurate conclusion can be obtained. We investigate this for the case of two stratigraphically/temporally ordered samples mentioned above and an assumed, yet realistic case of six ordered samples from the Hallstatt period.

The Two-Sample Case

To simplify matters we assume that there is no difference between the ^{14}C age and the calendar age. This corresponds to a strictly linear calibration curve. Sample *A* shall be measured as 1000 ± 100 AD (1σ), and *B* as 1030 ± 100 AD. Both probability distributions are modeled as Gaussian distributions to allow an analytical solution of the integrals involved, and the two distributions represent our *measured data.* From the variety of possible priors we focus on "exponential" priors of the form

$$P^{prior}(t_A, t_B) = \begin{cases} 0 & \text{if } t_A \le t_B \text{ (" forbidden" case)} \\ e^{-\frac{t_B - t_A}{\tau}} & \text{if } t_A > t_B \text{ ("allowed" case)} \end{cases} . \qquad (3)$$

This set of priors fulfills the criteria of a good class of priors for robust Bayesian analysis (as given by Berger 1994), including easy computational handling and easy interpretation. Figure 1 shows what happens if various exponential priors are applied. It is clearly visible that the selection of the prior strongly influences the posterior result.

If $\tau > 0$, τ is the *expectation value* for the age difference between the two samples *A* and *B* $\Delta t = t_B - t_A$, i.e. the average value we would obtain for a large number of similar sample pairs. Choosing a prior is therefore reduced to guessing the parameter τ, the average of Δt. The limit $\tau \to 0$ forces that the two samples actually stem from the same year (irrespective of their true age difference) and assigns the weighted average of the two measured ages to both samples.

The uniform prior is the special case where $\tau \to \infty$, i.e. this prior implicitly assumes that on the average the samples will be infinitely far apart. This paradoxical implication is obvious if one considers that this prior assigns the same probability to every age difference, and so an age difference between 10,000 and 20,000 years is 100 times as likely as an age difference between 0 and 100 years. This explains the tendency of shifting the time span between the two samples *A* and *B* towards larger age differences as shown in Steier and Rom (2000). The two-sample case supports the general observation that a prior which appears neutral from one point of view (same probability for every age difference) is often not neutral by other measures (infinite expected age difference). We ignore the case $\tau < 0$, since the respective priors favor large age differences even more than the uniform prior.

Using the whole set of exponential priors (3) we get 95%-HPD intervals for the older sample *A* ranging from [876 AD...1154 AD] $(\tau \to 0)$ to [786 AD...1116 AD] $\tau = \infty$. Here one may argue that our set of priors also contains distributions that are clearly unreasonable and cause an overestimation of the possible age range. As suggested by Berger (1994), we reduce the sensible range for τ, but this is not possible by mathematical means alone. In this point also robust Bayesian analysis

https://doi.org/10.1017/S0033822200038236 Published online by Cambridge University Press

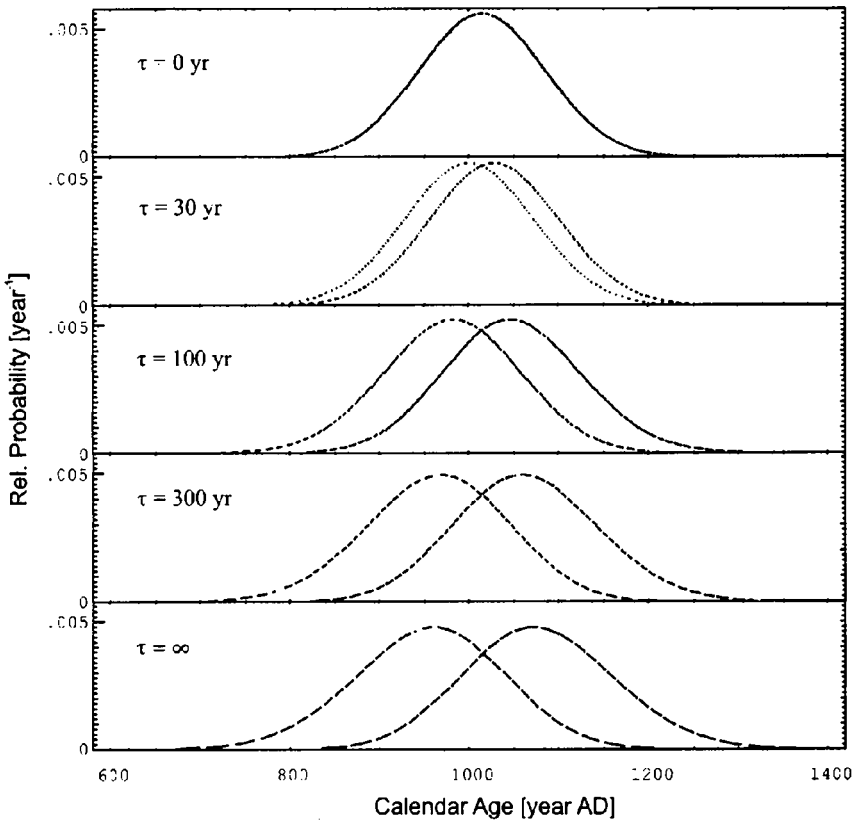

Figure 1 Marginal posterior probability distributions of a stratigraphically/temporally ordered pair of samples obtained for exponential priors with different values for the parameter τ (see text) assuming a strictly linear calibration curve. The probability distribution of the individually calibrated samples is assumed Gaussian with centers at 1000 AD and 1030 AD and a standard deviation of 100 years.

cannot completely avoid subjectivity. A constraint on τ is imposed by our knowledge that both samples are part of the same stratigraphy. But what is the general expectation value for the time span between different layers in a stratigraphy? An archaeologist may have a better guess, yet for our example we confine τ to the range between 30 and 1000 years.

We join all the resulting posterior 95%-HPD intervals for the confined range and obtain [788 AD...1144 AD] as the robust posterior 95%-HPD interval for sample *A*. If we compare this with the initial 95%-HPD interval for sample *A* [804 AD...1196 AD] where the stratigraphic information is not used, the robust posterior result appears plausible: both prior knowledge and measurement suggest that sample *A* is older than sample *B*, so we cannot learn much from sample *B* regarding the beginning of the 95%-HPD interval of *A*. On the other hand, the closest possible distance between the samples *A* and *B* is 0 years. In this case we can assign the mean age 1015 AD to both distributions, and the respective uncertainty is reduced by a factor of $\sqrt{2}$ i.e. we obtain the posterior interval for the exponential prior with $\tau \to 0$ (see above). It is easy to prove that no prior of type (2) yields a later end of the 95%-HPD interval for *A*. As a matter of course, our result is only valid if the set of priors we tried (exponential priors with 30 years $\leq \tau \leq$ 1000 years) forms a representative subset of the infinite number of possible priors.

https://doi.org/10.1017/S0033822200038236 Published online by Cambridge University Press

The Hallstatt Case

In this example we now increase the number of samples N and use the "real" INTCAL98 calibration curve (Stuiver et al. 1998) to study the influence of the respective wiggles. For a larger number N of temporally ordered samples, the prior $P^{prior}(t_1,..., t_N)$ of the (true) calendar ages $t_1,..., t_N$ is a multi-dimensional function with N arguments. The uniform prior

$$P^{prior}(t_1,...,t_N) = \begin{cases} const \quad for\ t_1,...,t_N\ in\ order & (\text{"}allowed\ case\text{"}) \\ 0 \qquad\qquad\qquad otherwise & (\text{"}forbidden\ case\text{"}) \end{cases} \tag{4}$$

shows a strong bias towards a larger *total time span* $t_N - t_1$ of the whole sequence (Bronk Ramsey 1999; Steier and Rom 2000). Please, note that there is a difference regarding the definition of the span between our papers (Steier and Rom 2000, and this paper) and Bronk Ramsey (1999).

In Steier and Rom (2000), we discussed a simulated set of six stratigraphically/temporally ordered samples from the Hallstatt period, i.e. the Early Iron Age in central Europe (750–400 BC), all samples having assumed true ages from the beginning of that period (see Figure 2). Due to the flatness of the calibration curve there, a simulated measurement yields almost identical ^{14}C ages for all six samples, and all 95%-HPD intervals obtained by individual calibration cover nearly the whole period. However, using the uniform prior we obtain posterior 95%-HPD intervals that disagree with the assumed true ages for the latest three samples, i.e. with one half of the whole sample set. In the present paper we now investigate whether robust Bayesian analysis is capable of solving the problem.

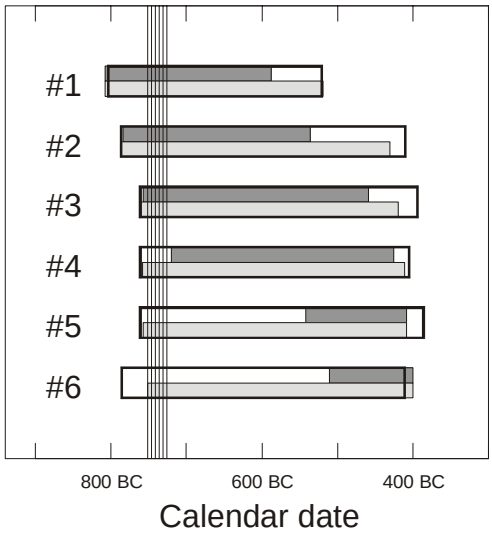

Figure 2 Robust Bayesian analysis applied to the simulated case of six ordered samples also discussed in Steier and Rom 2000. All true ages are assumed close to the beginning of the Hallstatt period (750–400 BC). Shown are the posterior 95%-HPD intervals of the individual calibrations (frames), Bayesian analysis with the uniform prior (see text, dark gray bars), and "robust Bayesian analysis" (see text, light gray bars). Small gaps in the 95%-HPDs were neglected for clarity.

Any realistic calibration curve is a "numerical" table, and the present case—as any other realistic case—cannot be treated analytically. Unfortunately, currently available calibration programs, which perform the integrations numerically, do not allow using freely defined prior probabilities. As a workaround, we explored only the existing possibilities supplied by OxCal to vary the prior (Bronk Ramsey 2000a): the "span-correcting prior" option was switched on and off, and "boundary" statements were added at various positions in the temporally ordered set of samples. We

https://doi.org/10.1017/S0033822200038236 Published online by Cambridge University Press

emphasize that we did not use this statement in the way intended by Bronk Ramsey, but we simply exploited its ability to modify the prior by placing the boundary statement in any position allowed by the program syntax. The priors involved in this workaround can be looked up in Bronk Ramsey (2000c). All resulting 95%-HPD intervals for each sample were then joined.

It is obvious that the set of priors that can be tested with this method is small and probably biased. However, even with these limitations, the joined posterior 95%-HPD intervals of the Hallstatt example after sensitivity testing are almost identical to the results of the single-sample calibration and cover the whole period. In our opinion, this is a plausible result: for a period where the calibration curve is flat, the ^{14}C measurement yields no useful information, and the ordering of the samples alone provides no means to assign the samples to a certain part of that period. We emphasize that none of the single priors by itself is capable of yielding this result (including the prior obtained by following the instructions in the latest versions of the OxCal program, which performs much better than the uniform prior).

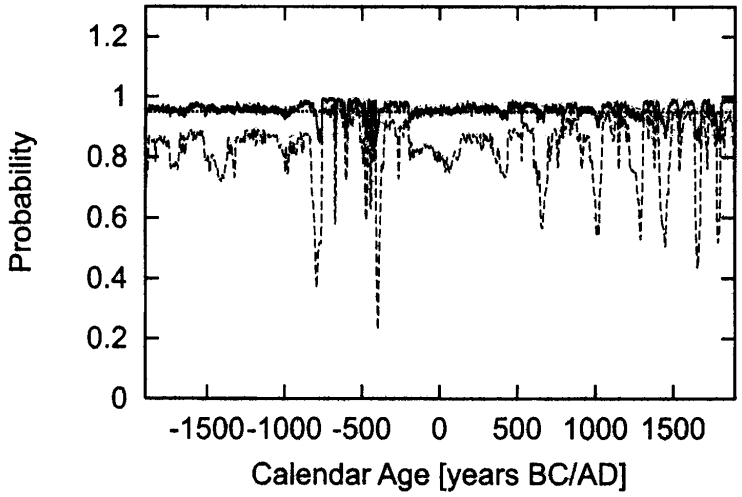

Figure 3 Samples originally distributed uniformly with regard to their true ages show uneven distributions of the respective single-calibrated ages. 1000 samples were simulated for every year, assuming a measurement uncertainty of 50 years. The portion of samples that lie inside their 95%-HPD interval (solid curve) and the sum probabilities of all samples (dashed curve) are shown. These structures, induced by the wiggles in the calibration curve, are an artifact of the method and are of no archaeological significance. The dotted horizontal line is the 95% level.

A Subtlety in Single-Sample Calibration

In ^{14}C calibration by means of Bayesian mathematics there exists a subtlety in the definition of the 95%-HPD intervals. These intervals are designed such as to contain the correct (true) age for 95% of all calibrated samples if their (true) age distribution matches the prior distribution used for the calibration. In the case of single sample ^{14}C calibration this implies that one cannot expect to obtain the correct result for 95% of all the samples originating in a particular year (since generally a uniform prior is used, which regards each true age equally likely). The yield of correct results will vary from year to year, and only on the average of all the years that are covered by the calibration curve one will get 95% (Bronk Ramsey 2000b). We studied this subtlety by computer-simulated measurements using the uniform prior $P^{prior}(t_A)=const.$ (see Figure 3). From this, samples originally distributed uniformly with regard to their true ages show an uneven distribution of the respective single-

https://doi.org/10.1017/S0033822200038236 Published online by Cambridge University Press

calibrated ages. These structures, induced by the wiggles in the calibration curve, are an artifact of the method and are of no archaeological significance.

For an assumed large set of samples originating from a stable culture, which covers the centuries around the beginning of the Hallstatt period (750 BC), Figure 3 shows that the years around 800 BC will be included in exceptionally few single-calibrated 95%-HPD intervals. However, a large number of samples is required to produce a significant effect in a data set, so that a possible misinterpretation seems unlikely. Much more distinct minimums show up in the so-called *sum probability*, which is obtained by averaging the single-calibrated probability distributions. Besides other peculiarities (see Bronk Ramsey 2000c), this is another reason to be cautious in the interpretation of sum calibrations. Such a structure may be erroneously interpreted as a hint for a true discontinuity in a culture.

SUMMARY AND CONCLUSIONS

A critical point in Bayesian statistics is the conversion of qualitative archaeological information into mathematical prior probabilities. Since in many cases assumptions/guesses are needed to construct these probability functions, it is important to assure that these assumptions/guesses do not bias the results. By varying these assumptions (to reflect uncertain and incomplete prior information or knowledge) and subsequently using only results that are robust under these changes this can be achieved.

Obviously, our method to perform this "sensitivity testing" by exploiting OxCal's "boundary" statement in a way that was not intended by the programmer does not allow testing an adequate set of priors. The actual application of robust Bayesian analysis for real archaeological problems would require a computer code that allows defining the prior freely but which so far is not commonly available.

However, even with these limitations, the following preliminary conclusions are possible: the increase in precision (i.e. smaller 95%-HPD intervals) that can be achieved with Bayesian statistics is much smaller than promised by the uniform prior, which was commonly used in archaeology before 1999. Despite the difficulty to find one general-purpose prior, testing a range of priors will yield a joined result which is probably not more precise, but which is more accurate than the ^{14}C intervals obtained from using any individual prior (i.e. a smaller part of the true sample ages will lie outside the quoted 95%-HPD intervals).

REFERENCES

Bayes T. 1763. An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London* 53:370–418. A Postscript and a LaTeχ source file version of this paper are available at URL http://www.york.ac.uk/depts/maths/histstat/essay.ps, and http://www.york.ac.uk/depts/maths/histstat/essay.htm, respectively.

Berger JO. 1980. *Statistical decision theory.* New York: Springer-Verlag. 425 p.

Berger JO. 1994. An overview of robust Bayesian analysis. *Test* 3:5–124.

Bronk Ramsey C. 1999. An introduction to the use of Bayesian statistics in the interpretation of radiocarbon dates. In: *Proceedings of the International Workshop on Frontiers in Accelerator Mass Spectrometry.* 6–8 January 1999. National Institute for Environmental Studies, Tsukuba. National Museum of Japanese History, Sakura. Japan: p 151–60.

Bronk Ramsey C. 2000a. Comment on 'The use of Bayesian statistics for ^{14}C dates of chronologically ordered samples: a critical analysis'. *Radiocarbon* 42(2):199–202.

Bronk Ramsey C. 2000b. OxCal Program v3.5 Manual. URL: http://www.rlaha.ox.ac.uk/orau.

Bronk Ramsey C. 2001. Development of the radiocarbon calibration program OxCal. *Radiocarbon*. This issue.

https://doi.org/10.1017/S0033822200038236 Published online by Cambridge University Press

Bronk Ramsey C, van der Plicht J, Weninger B. 2001. "Wiggle-matching" radiocarbon dates. *Radiocarbon*. This issue.

Buck CE, Cavanagh WG, Litton CD. 1996. Bayesian approach to interpreting archaeological data. Chichester, New York, Brisbane, Toronto, Tokyo, Singapore: John Wiley & Sons. 382 p.

Jeffreys H. 1961. *Theory of probability.* 3rd edition. Oxford: Clarendon Press. 447 p.

Malakoff D. 1999. Bayes offers a 'new' way to make sense of numbers. *Science* 286:1460–4.

Steier P, Rom W. 2000. The use of Bayesian statistics for ^{14}C dates of chronologically ordered samples: a critical analysis. *Radiocarbon* 42(2):183–98.

Stuiver M, Reimer PJ, Bard E, Beck JW, Burr GS, Hughen K, Kromer B, McCormac G, van der Plicht J, Spurk M. 1998. INTCAL98 radiocarbon age calibration, 24,000-0 cal BP. *Radiocarbon* 40(3):1041–83.

Yang R, Berger JO. 1997. A catalog of noninformative priors. *Institute of Statistics and Decision Sciences: Working Paper Series*. URL: http://ftp.isds.duke.edu/WorkingPapers/97-42.html. Accessed 18 October 2000.