Rapid Communication

All age–depth models are wrong: but how badly?

R.J. Telford\textsuperscript{a,}\textsuperscript{*}, E. Heegaard\textsuperscript{a,b}, H.J.B. Birks\textsuperscript{a,b,c}

\textsuperscript{a} Bjerknes Centre for Climate Research, University of Bergen, Allégaten 55, N-5007 Bergen, Norway
\textsuperscript{b} Botanical Institute, University of Bergen, Allégaten 41, N-5007 Bergen, Norway
\textsuperscript{c} Environmental Change Research Centre, University College London, 26 Bedford Way, London WC1H 0AP, UK

Received 22 September 2003; accepted 6 November 2003

Abstract

The uncertainty and error associated with fitted age–depth models were investigated by fitting a range of models to sets of simulated radiocarbon dates based on varved sediments and comparing the modelled relationship with the varve timescale. With large numbers of dates, cubic smooth splines out-perform other models; with few dates, no model reliably provides a good fit to reality, and calculated confidence intervals are over-optimistic.

\textsuperscript{*}Corresponding author.
E-mail address: richard.telford@bjerknes.uib.no (R.J. Telford).

\textsuperscript{0277-3791/$-see front matter \textcopyright 2003 Elsevier Ltd. All rights reserved.

doi:10.1016/j.quascirev.2003.11.003

1. Introduction

An absolute chronology is the basis of all comparisons and correlations of Late Quaternary stratigraphic proxy records. With the exception of varved (annually laminated) sediments, the chronology of most lacustrine and marine sediment cores is derived from radiocarbon dates. The construction of an age–depth relationship from these dates is a non-trivial task. Many procedures for constructing age–depth relationships have been suggested. Commonly used models are linear interpolation, splines, and linear regression models (Bennett, 1994). Other models include Bézier curves (Bennett and Fuller, 2002), mixed-effect models (Heegaard, 2003) and fuzzy regression (Boreux et al., 1997). Unfortunately, the different approaches can give very different answers (Bennett, 1994; Bennett and Fuller, 2002), both for the age at a particular depth and the uncertainty attached to that estimate, and it is difficult to determine which relationship is closest to reality.

Age–depth models are only meaningful and useful when calculated using calibrated radiocarbon dates (Bartlein et al., 1995): age–depth models calculated on uncalibrated radiocarbon dates make the implicit, and implausible, assumption that variations in the sedimentation rate cancel out the wiggles in the calibration curve. Therefore, all the models used in this study are based on calibrated dates. Unfortunately, calibration of radiocarbon dates adds an extra level of complexity (Bennett, 1994) as the resulting probability distributions are not Gaussian.

In this study, we used simulated radiocarbon dates uniformly spaced along varved sediment sequences (Fig. 1). These radiocarbon dates, with a stipulated error, are then calibrated and different age–depth models constructed. A comparison between the predicted and true age of each varve can then be made. This study tests: (1) whether one model reliably out-performs the others with large and small numbers of dates; (2) if the uncertainties estimated by the model are realistic; and (3) if any diagnostic statistics are a useful guide to model choice.

Brauer et al. (2001) report that one of the varved sequences used in this paper may contain errors, but, provided that the changes in sedimentation rate implied by the apparent varve chronology are realistic, this does not affect the results of the methodology employed here.

2. Method

Two varved sequences (Fig. 1), Holzmaar (Zolitschka et al., 1998) and Elk Lake, Minnesota (Bradbury and Dean, 1993) were used in the analysis. Both sequences are 10–12,000 years long.

Between six and 48 dates were equally spaced along each core and de-calibrated with the INTCAL98
calibration curve (Stuiver et al., 1998) to give radiocarbon years (See Fig. 2). Radiocarbon dates always have an uncertainty associated with them, typically 30–100 years for Holocene sediment; this was modelled by adding a single normal deviate from a distribution with zero mean and standard deviation of 40 years. Dates were calibrated with winBUGS (Spiegelhalter et al., 2003) using models modified from Millard (2003) that use the a priori information that the dates are in stratigraphic order (Buck et al., 1996). The initial values for the winBUGS analysis were drawn from a normal distribution, with a standard deviation of 40 years, and a mean equal to the intercept of the radiocarbon date with the calibration curve. The weighted average of the probability distribution of the calibrated date is used as a midpoint for model construction.

3. Results

After de-calibration and re-calibration, the proportion of dates that have high-probability ranges that encompass their true age is 69% and 95% at 1 and 2σ, respectively. Plots of the difference between predicted age and varve age against varve age for one set of nine simulated dates from Holzmaar are shown in Fig. 3. Although the dates used have an error of ±40 years, the predicted dates are up to 400 years out, and the RMSEP varies between 98 years for the spline interpolation model and 156 years for the fifth-order polynomial model. None of the confidence intervals contained 68% of the true ages of the varves, as would be expected if they were correct; the fifth-order polynomial model confidence interval contained only 21% of the true varve ages.

Fig. 4 shows that as the number of dates used increases, RMSEP declines, and the proportion of varves with ages within the confidence interval generally...
increases. These results are the median of 10 trials. Linear interpolation models have the lowest RMSEP, with few dates, at Holzmaar, but the largest RMSEP at Elk Lake. At both sites, with 48 dates, smooth cubic splines have the lowest RMSEP, approaching the uncertainty of the dates, whereas polynomials have the highest RMSEP. In both lakes, the proportion of varves within the confidence intervals is low when there are few dates (both at 1 and 2σ: the latter is not shown). With a large number of dates the confidence intervals become more realistic, except for polynomial models, which are always, especially at Holzmaar, too optimistic.

4. Discussion

The use of simulated radiocarbon dates on varved sediments is a novel approach to test the validity of age–depth models. Boreux et al. (1997) use a sediment sequence with a large number of dates to test their fuzzy age–depth models. They fitted models using a subset of the dates, and used the remainder to validate it. It is difficult to quantify either the bias or the error in this approach, other than the fits are reasonable. Bennett and Fuller (2002) fit a variety of age–depth models to see which predicts the age of the hemlock decline in eastern North America nearest to its estimated age. They find that linear interpolation is “hard to beat” with the number of dates available from the sites they used. Unfortunately, the date of the Hemlock decline is weakly constrained, at best to ± 50 14C years; it provides only a single point at which to check the age–depth model; and the validity of confidence intervals cannot be checked.

The choice of the number of degrees-of-freedom, and of the polynomial order, for smooth splines and weighted linear regression has a large impact on the resulting age–depth model (see Bennett and Fuller, 2002). If too low a polynomial order, or number of degrees-of-freedom, is used, the model is too stiff; too high and the model can show reversals as it fits noise or becomes numerically unstable. Therefore, diagnostics that aid model choice are very important: generalised cross-validation, chi-squared goodness-of-fit, and $r^2$ have all been used. The latter is not a useful guide: provided that the sediment is in stratigraphic order and deposited over a long period of time, $r^2$ will always be high, and increase with higher-order polynomials.

Chi-squared goodness-of-fit and generalised cross-validation, for linear regression and smooth spline models, respectively, normally select a near-optimal number of degrees-of-freedom, as assessed by RMSEP, although even the optimal model can still be a poor representation of reality. Smooth splines generally use more degrees-of-freedom than selected for linear regression, probably because they do not suffer from numeric instability with many degrees-of-freedom. This is especially so for analyses with few dates, where smooth splines and polynomials can be too stiff: a poor chi-squared goodness-of-fit can provide a warning that the model is unlikely to be true given the radiocarbon dates.

That the confidence intervals produced by all the models were unrealistically narrow for less than 12 dates (more in Elk Lake) is worrying, as this is the typical number of dates on a well-dated Holocene sequence, and was initially surprising. On reflection, the models will only generate realistic confidence intervals if their assumptions are not violated. For example, the confidence intervals on a fifth-order polynomial model have the assumption that the sediment deposition followed a fifth-order polynomial, and a linear interpolation model assumes that sedimentation rates change abruptly at the depths of the dates. When a model is fitted to a large number of dates, these assumptions will be approximately true, but not necessarily so for just a few dates.
Bennett (1994) promotes the use of linear regression models, as they have narrow confidence intervals. Narrow confidence intervals are desirable, but only if they are realistic.

While it does not appear to be possible to predict which age–depth model will give the best fit to reality, some general points can be made. Linear regression models have the undesirable property that they are a global fit to the data; changes at one end of the core can affect the relationship fitted to the whole stratigraphy. Linear and spline interpolation force the age–depth model to pass through the dates. This means that the models cannot deviate too far from reality, in contrast to smoothing splines and linear regression models that can bypass a date (Fig 3a). Since the dates, given the uncertainties of radiocarbon dating, are unlikely to be the true date of that sediment, forcing the model to pass through every date incorporates this noise into the model. If the density of dates is high relative to their uncertainty, smoothingsplines can attempt to ignore this noise and result in a better fit. In this study, this appears to happen when there are approximately 24 dates in a Holocene length sequence: few real sequences have this many dates.

This study may be unduly pessimistic when small numbers of dates are used, as spacing dates equidistant down core may not be the optimal strategy. If changes in the sedimentation rate can be identified, either from sediment lithology or by selecting samples to date in an iterative fashion, a more precise chronology will be developed, and interpolation methods most appropriate. However, in all other aspects, the age–depth relationships in this study are constructed under optimal conditions: all the simulated dates are accurate, and their true uncertainty is known; the dates are taken on a single year’s sediment accumulation; and there are no hiatuses. This contrasts with real datasets where dates can suffer from contamination with old and modern carbon, reservoir effects; possible under- or over-estimation of the dates’ precision (Boaretto et al., 2002); the dated sediment sample represents an unknown number of years with a possibly unequal rate of carbon accumulation; and unrecognised hiatuses can span decades or centuries (Brauer et al., 2001). Therefore, the model accuracy and validity of the errors shown here may be better than what can be achieved in practice.

All the age–depth models considered above are constructed using the weighted average of the calibrated radiocarbon dates’ probability distribution. The use of a central point estimate in this way is usual practice, but not entirely appropriate, especially when the radiocarbon date falls into one of the radiocarbon plateaux. Here the wide and multimodal probability distributions cannot be meaningfully characterised by a single value, and large errors may result if this is done. One approach that uses the full probability distribution is wiggle matching. Wiggle-matching age–depth models, which

Fig. 4. Median root mean square error of prediction (left), and proportion of varves with their true age within the modelled 1σ confidence interval, and the expected value of 0.68 (horizontal line), (right) against number of dates for (a) Holzmaar, and (b) Elk Lake. The values are the median of 10 trials.
calibrate the radiocarbon dates using the prior belief that the sedimentation rate is linear, or piece-wise linear, using either a Bayesian (Christen et al., 1995) or least squares approach (Blaauw et al., 2003), will produce narrower model confidence intervals than shown here. However the behaviour of these models when the assumption of constant sedimentation is violated has not been investigated: it is likely that as the deviation from linearity becomes larger than the standard deviation of the dates, results will become increasingly inaccurate, with precision spuriously high. Both Elk Lake and Holzmaar varve sequences have few sections longer than 1000 years that meet even this weak definition of linearity, so very many radiocarbon dates would be needed to characterise and date these sequences with wiggle matching.

5. Conclusions

The construction of age–depth relationships from simulated radiocarbon dates on varved sediments is a powerful tool to test the validity of age–depth modelling techniques. It has shown that uncertainties calculated from the error on the radiocarbon dates underestimate the true uncertainty of the age–depth model, especially when there are few dates, and that no model reliably performs well when there are only a few dates. Smooth cubic spline models have the closest fit to reality when there are many dates relative to changes in sediment accumulation rate. If changes in sedimentation rate can successfully be identified from sedimentological or other evidence, and dated, then interpolation models are probably the most appropriate.

The number of dates needed to construct an age–depth model for a sediment sequence will depend on the required precision and the complexity of the sedimentation rate. It is probable that to achieve accurate, high-precision, radiocarbon-based chronologies, researchers will have to use many more dates than is currently the norm, and uncertainties may always be high during radiocarbon plateaux. The ages of events are probably better constrained by dating them directly, ideally with dates immediately, but unambiguously, above and below the event horizon, than by age–depth models. As chronological uncertainty will increase when comparing sites, full use should be made of time parallel markers, such as tephra layers and geomagnetic stratigraphy.

Acknowledgements

We thank Andrew Millard for useful discussion and help with winBUGS and Simen Blockley who reviewed the paper. This research was supported by NORPAST-2, funded by the Norwegian Research Council, and PACLIWA, an EU Framework 5 Programme project (EVK2-CT2002-000143). This is publication No. A28 of the Bjerknes Centre for Climate Research.

References


Telford, R.J., Heegaard, E., Birks, H.J.B., in press. The intercept is a poor estimate of a calibrated radiocarbon age. The Holocene. in press.