RADIOCARBON, Vol 42, Nr 2, 2000, p 183–198

THE USE OF BAYESIAN STATISTICS FOR ^{14}C DATES OF CHRONOLOGICALLY ORDERED SAMPLES: A CRITICAL ANALYSIS

Peter Steier • Werner Rom

Vienna Environmental Research Accelerator, Institut für Radiumforschung und Kernphysik, Universität Wien, Währinger Strasse 17, A-1090 Vienna, Austria. Email: peter.steier@univie.ac.at.

ABSTRACT. Bayesian mathematics provides a tool for combining radiocarbon dating results on findings from an archaeological context with independent archaeological information such as the chronological order, which may be inferred from stratigraphy. The goal is to arrive at both a more precise and a more accurate date. However, by means of simulated measurements we will show that specific assumptions about *prior probabilities*—implemented in calibration programs and not evident to the user—may create artifacts. This may result in dates with higher precision but lower accuracy, and which are no longer in agreement with the true ages of the findings.

INTRODUCTION

In many cases, the radiocarbon age is not the only information available on archaeological samples. Additional information may originate from typology, stratigraphy, or dendrochronology. Whereas ^{14}C measurements directly provide probability distributions (due to the inherent Poisson statistics of the counting process), typology and stratigraphy do not. In a mathematical sense, they give non-probabilistic logical statements such as "event A is earlier than event B" or "object A typologically matches object B". The classical statistical approach tends to reduce the ^{14}C distributions also to logical statements like "the age of the sample lies between 3360 BC and 3100 BC" using 95% confidence intervals. This is then combined with the additional archaeological evidence by means of scientific reasoning (see Reece 1994).

As an alternative, the additional archaeological information may also be transformed into probability distributions. All the information may then be integrated by using Bayesian mathematics (for an overview see Litton and Buck 1995; Buck et al. 1996; for applications see Buck et al. 1991, 1992, 1994; Bayliss et al. 1997). The additional archaeological information investigated in this paper is the chronological order of the samples. In this case, the intention behind applying the Bayesian method is to improve the date obtained from the ^{14}C measurement alone. Since the knowledge of the chronological sequence adds independent information, this appears feasible. However, we will show through computer-simulated measurements that assumptions used to transform the additional archaeological information into probability distributions may create results with higher precision (i.e. reduced uncertainties) of dates, but lower accuracy (i.e. reduced agreement with the true ages of the samples).

THE BAYES ALGORITHM (BAYES' THEOREM)

In evaluating experimental data the so-called *Bayes' theorem* (Bayes 1763) plays a fundamental role. Bayes' theorem allows us to combine measured data from a sample with our knowledge on the corresponding sample before (*prior* to) the measurement. Both the measured data and the prior information must be formulated mathematically as probability distributions. After feeding them into Bayes' theorem we get the so-called *posterior* probability distribution which incorporates both measured and prior information. Since the main features in applying Bayes' theorem already show up in the ^{14}C dating of single samples, we will discuss this case first. It will provide useful results needed for the subsequent investigation of the multiple sample case, and shall also serve as an illustration of Bayes' theorem. For readers not familiar with the mathematics involved in ^{14}C calibration, we suggest Buck (1996:203–15).

Calibration of a Single Sample

The data collected in a ^{14}C measurement are reduced to the ^{14}C age t^{C14} and its uncertainty σ. We neglect the asymmetry of the uncertainty, which is induced by the exponential shape of the decay curve, and is only significant for very old samples. From the ^{14}C age we want to derive the true age of the sample on the calendar age scale. For a single sample the procedure is the usual ^{14}C calibration process. We try to look up the age t^{cal}, i.e. the calibrated or calendar age, where the ^{14}C age from the tree-ring calibration curve C(t) matches the ^{14}C age t^{C14} of the sample

$$t^{cal} = C^{-1}(t^{C14}) \tag{1}$$

with C^{-1} being the mathematical inversion of the calibration curve. Unfortunately, in the general case the calibration curve is not an invertible mathematical function. Due to its "wiggles" one can get more than one match, and its uncertainty also complicates the situation. Bayes' theorem is the mathematical tool suited to invert the calibration function in a probabilistic sense. We use it to get the probability that the sample has a certain calendar age t with respect to the measured ^{14}C age t^{C14}. Let us denote this probability as $P^{cal}(t)$. $P^{cal}(t)$ is the *posterior* probability for the ^{14}C calibration of a single sample.

A statistical model of the underlying measurement and calibration process allows us to determine the probability of how likely an (assumed) calendar age t for the sample of interest is going to yield the data t^{C14} observed in the actual measurement. Bayesian mathematics calls this probability distribution the *likelihood function* $P^{likelihood}(t^{C14}|t)$. The likelihood function for the calibration of a single ^{14}C date is

$$P^{likelihood}(t^{C14}|t) = \frac{1}{U}\frac{1}{\sqrt{2\pi\bar{\sigma}^2}} e^{-\frac{(t^{C14}-C(t))^2}{2\bar{\sigma}^2}} \tag{2}$$

where $\bar{\sigma}^2 = \sigma^2\left[t^{C14}\right] + \sigma^2\left[C(t)\right]$ with

$\sigma^2\left[t^{C14}\right]$ . . . . . uncertainty of the ^{14}C age t^{C14} from the measurement

$\sigma^2\left[C(t)\right]$ . . . . . uncertainty of the calibration curve at the assumed true age t.

U . . . . . a normalization constant to achieve $\int_{-\infty}^{\infty} P^{likelihood}(t^{C14}|t)\,dt = 1$.

Since both $C(t)$ and $\sigma^2\left[C(t)\right]$ heavily depend on t, the likelihood function is not Gaussian in shape unless the assumed calibration curve is strictly linear with constant uncertainty.

The difference between $P^{cal}(t)$ and $P^{likelihood}(t^{C14}|t)$ is essential, since one needs Bayes' theorem to derive the latter probability from the former. The formulation of Bayes' theorem to calibrate a single ^{14}C date is

$$P^{cal}(t) = \frac{1}{U'} P^{likelihood}(t^{C14}|t) \cdot P^{prior}(t) \quad . \tag{3}$$

U´ is a constant needed to normalize $\int_{-\infty}^{\infty} P^{cal}(t)\,dt$ to unity.

The only unknown in formula (3) is $P^{prior}(t)$, the probability distribution of the true age prior to the measurement. Bayes' theorem is (implicitly) used for a variety of problems. In most cases, the prior probability is not known exactly. This also holds for the tree-ring calibration of a single ^{14}C date. However, in this case the likelihood function (2) disappears sufficiently fast outside a relatively small region (we neglect cases where the ^{14}C age is consistent with infinity). The assumption that the prior probability $P^{prior}(t)$ is approximately constant and different from zero in this region is sufficient to apply Bayes' theorem and we obtain

$$P^{cal}(t) \equiv P^{likelihood}(t^{C14} \mid t) \; . \tag{4}$$

Since $P^{likelihood}(t^{C14} \mid t)$ is already normalized to unity, the constant U is no longer needed, and the posterior probability is identical to the likelihood function in this easy case. $P^{cal}(t)$ is the function usually plotted on the calendar age scale of the calibration diagrams (see Figure 5). For every "wiggle" of the calibration curve that crosses or touches the ^{14}C age t^{C14} of the sample we get a local maximum in $P^{cal}(t)$.

The posterior probability distribution $P^{cal}(t)$ is reduced to 95% confidence intervals: after tabulating calendar years and their corresponding probabilities a set of years is accumulated until a total probability of 95% is reached. A small degree of freedom remains in terms of which years to collect first (Buck 1996:152–3), but all resulting intervals share the feature that 95% of the true ages of the samples (should) lie inside, and 5% (should) lie outside. In this paper we collect the years with the highest probabilities first, as usual in archaeological dating. Local maxima in the probability density may lead to several disjunct intervals.

Multiple Samples

In a more general formulation of Bayes' theorem, the true values of a set of parameters and the corresponding measured values shall be denoted as *true values* and *measured data*, respectively. Formula (3) then reads as

$$\begin{aligned} &P^{posterior}(\text{true values} \mid \text{measured data}) = \\ &= \frac{1}{U'} P^{likelihood}(\text{measured data} \mid \text{true values}) \cdot P^{prior}(\text{true values}) \quad . \end{aligned} \tag{5}$$

If the prior probability P^{prior} (true values) is sufficiently constant where the likelihood function is not zero, then the posterior probability distribution $P^{posterior}$ is identical with $P^{likelihood}$.

Compared to the calibration of a single date, the situation is more complicated for the combination of the ^{14}C dating results of N independently measured samples. Every sample has an (unknown) calendar age and a measured ^{14}C age denoted by t_k and t_k^{C14}, respectively, for the sample with index k =1, ...,N. The additional information included in the statement "the chronological order of the samples is 1, 2, 3,..." can be transformed into this common N-dimensional prior probability:

$$P^{prior}(t_1, t_2, ..., t_N) = \begin{cases} \text{const for } t_1, ..., t_N \text{ in order} & (\text{"allowed case"}) \\ 0 \quad \text{otherwise} & (\text{"forbidden case"}) \end{cases} \; . \tag{6}$$

The ^{14}C measurement yields N probability distributions for the calibrated ^{14}C ages, $P_k^{cal}(t_k)$, which are the posterior probabilities of the single-sample calibration. They are now combined to a N-dimensional likelihood function for a second application of Bayes' theorem:

$$P^{likelihood}(t_1^{C14},...,t_N^{C14} \mid t_1,...,t_N) = \frac{1}{U} P_1^{cal}(t_1) \cdot ... \cdot P_N^{cal}(t_N) . \tag{7}$$

This fulfills the definition of $P^{likelihood}$ in the general case of Bayes' theorem in (5) because of (4) and the independence of the single sample likelihood functions in a probabilistic sense. We neglect the complex correlations induced by the uncertainty of the calibration curve (see Buck 1996:235–7).

Next, all the information is combined to get the posterior probability distribution

$$P^{posterior}(t_1,...,t_N) = \frac{1}{U'} P^{likelihood}(t_1^{C14},...,t_N^{C14} \mid t_1,...,t_N) \cdot P^{prior}(t_1,...,t_N) \quad . \tag{8}$$

The so-called *marginal* posterior probability distribution for the single samples are obtained by integrating over all possible dates t_k of the respective other samples. Using the definitions given above we get

$$P_k^{posterior}(t_k) = \frac{1}{U'_k} \int_{-\infty}^{\infty} dt_1 \int_{-\infty}^{\infty} dt_2 ... \int_{-\infty}^{\infty} dt_{k-1} \int_{-\infty}^{\infty} dt_{k+1} ... \int_{-\infty}^{\infty} dt_N P_1^{cal}(t_1) \cdot ... \cdot P_N^{cal}(t_N) \cdot P^{prior}(t_1,t_2,...,t_N) =$$

$$\frac{1}{U'_k} \int_{-\infty}^{t_k} dt_1 \int_{t_1}^{t_k} dt_2 ... \int_{t_{k-2}}^{t_k} dt_{k-1} \int_{t_k}^{\infty} dt_{k+1} \int_{t_{k+1}}^{\infty} ... \int_{t_{N-1}}^{\infty} dt_N P_1^{cal}(t_1) \cdot ... \cdot P_N^{cal}(t_N) \tag{9}$$

where the U'_k denote constants needed to normalize $\int_{-\infty}^{\infty} P_k^{posterior}(t_k) dt_k$ to unity.

In this paper we will call the method of combining sample ordering information and ^{14}C data the "sequence algorithm". Our analytical formulation using a multidimensional integration is equivalent to the Monte Carlo method ("Gibbs sampling") presented in Buck et al. (1992).

As we will show, problems in the sequence algorithm arise from the assumption that $P^{prior}(t_1, t_2, ..., t_N)$ is constant in the "allowed case" (see the common prior [6]). The resulting marginal posterior probabilities $P_k^{posterior}(t_k)$ are highly dependent on this assumption in regions where the ^{14}C likelihood functions $P_k^{prior}(t_k^{C14} \mid t_k)$ do not disappear (see e.g. Roe 1992; Buck 1996:170–1; Blobel and Lohrmann 1998). By means of simulated measurements we investigated the consequences of applying this algorithm.

THE SEQUENCE ALGORITHM APPLIED TO COMPUTER-SIMULATED ^{14}C MEASUREMENTS

The most persuasive test for the sequence algorithm would be a set of real samples with known true ages from the same archaeological context. The algorithm should then be applied to the ^{14}C data, and the resulting dates could be compared with the true ages. Although measurements on such data

sets may have been performed in the past (we know of none), a large number is required for a thorough check of the algorithm. Therefore we used artificial data sets on which we performed computer-simulated measurements.

In the mathematical analysis given above we incorporated the calibration process into the likelihood function, using calendar ages for the true values but ^{14}C ages as the measured data (see equation [5]). For our further investigations the details of the single-sample calibration are not essential. Therefore we consider the resulting probability densities $P_k^{cal}(t_k)$ on the calibrated age axis as the measured data. We apply the sequence algorithm to simplified sets of $P_k^{cal}(t_k)$ suitable to study the separate influence of various parameters on the posterior results. If not otherwise mentioned, we will use Gaussian-shaped calibrated age distributions $P_k^{cal}(t_k)$ since this allows us to solve (9) analytically. For non-Gaussian distributions we use the computer program OxCal v2.18 (Bronk Ramsey 1995a, 1995b) which implements the "Gibbs sampling" method mentioned above. Due to the different features of different parts of the calibration curve there are two extreme cases:

The "linear" case: In some regions the calibration curve can be approximated by a strictly linear function without any wiggles. Since we assumed the probability distribution for the ^{14}C age to be Gaussian-shaped, in this case the calibrated probability densities $P_k^{cal}(t_k)$ will be roughly Gaussian-shaped also. In addition we assume that the uncertainty of the calibration curve is negligible compared to the uncertainty of the ^{14}C data. If several samples with the same true age are independently ^{14}C-dated, then the scatter of the centers of the $P_k^{cal}(t_k)$ should match their width in this case.

The "flat" case: In other regions the calibration curve is flat and largely dominated by wiggles. The $P_k^{cal}(t_k)$ are not Gaussian-shaped, and they span from the first to the last crossing (or proximity) of the calibration curve and the measured ^{14}C age. Since this is due to the features of the calibration curve and not due to the ^{14}C measurement uncertainty, the 95% confidence intervals are essentially the same for all samples. Large uncertainties in the calibration curve have a similar effect. The real ^{14}C calibration curve is somewhere in between these two extreme cases.

Modeling without statistical scatter is much easier since for every simulation only one set of input data exists. If scatter is included the simulation has to be performed with a sufficiently large number of randomly generated data sets to get a significant result. Therefore, in most simulations the scatter is ignored. In this case we check for selected points whether the qualitative result is influenced by scatter. However, the "flat" case shows that simulations without scatter have a value on their own.

Computer Experiment A

In this computer experiment we investigate six ^{14}C samples within a chronological sequence. The calibrated ^{14}C data were modeled by Gaussian probability distributions with a standard deviation σ of 100 yr. Every data set consisted of six samples with constant time spacing Δt. We used sets with Δt of 0, 1, 2, 5, 10, 20, 25, 50, 75, 100, 125, 150, 175, 200, 300, 400, 500, 750, and 1000 yr. We model the calibrated ^{14}C probability distributions with their centers exactly at the true ages t_k^{true}, so we ignore any statistical scatter (see Figure 1):

$$t_1^{true} = 1000\ BC - \frac{N-1}{2}\,\Delta t \quad ; \quad t_k^{true} = t_1^{true} + (k-1)\,\Delta t \ \text{ with } \ k = 2,\ldots,N \tag{10}$$

$$P_k^{cal}(t_k) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(t_k - t_k^{true})^2}{2\sigma^2}} \quad . \tag{11}$$

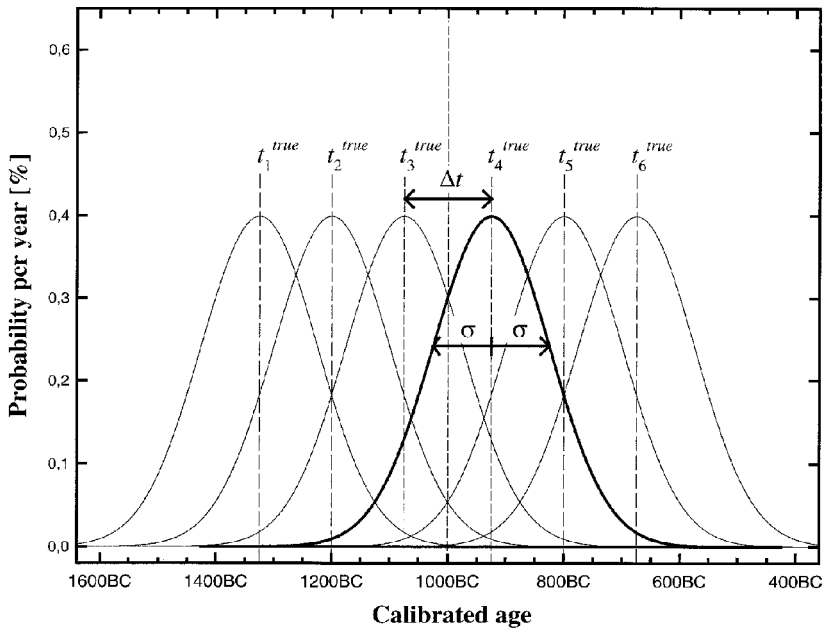

Figure 1 Computer experiment A. The Bayesian sequence algorithm is applied to sets of 6 ordered samples. Each set is constructed symmetrically around 1000 BC with constant time spacing Δt between the true ages t_k^{true}. The individually calibrated probability distributions before applying the sequence algorithm are assumed to be Gaussian-shaped with their centers exactly at t_k^{true} and with σ = 100 yr.

The influence on the data for sets with different Δt is shown in Figure 2, where we focus on the youngest (latest) sample (#6). We compare the artificial ^{14}C data and the posterior data resulting from the sequence algorithm.

It turns out that for Δt considerably larger than 1σ (100 yr), the data and the corresponding uncertainties are not modified significantly. For short Δt the algorithm shifts apart the probability distributions to cover the whole interval compatible with the ^{14}C measurement uncertainty. In this region the posterior uncertainty is reduced, i.e. precision increased. This is the case for which the sequence algorithm was developed in the first place.

From Figure 2 one can see that the algorithm shifts the age of the latest sample (#6) towards the assumed measurement uncertainty. Near Δt = 0 yr the result is independent of the true ages, but is determined by the measurement error! The probability distributions are no longer in agreement with the assumed true ages, so in our opinion the increased precision is an artifact.

We want to complement our investigation and verify that the kind of statistical scatter which shows up in the previously mentioned "linear" case does not influence the qualitative result. We choose the case with Δt = 0 yr (all samples exactly from 1000 BC). The calibrated ^{14}C data are modeled as above, but now we allow random shifts of the centers of the probability distributions:

$$P_k^{cal}(t_k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(t_k - t_k^{true} - \xi_k)^2}{2\sigma^2}} \qquad (12)$$

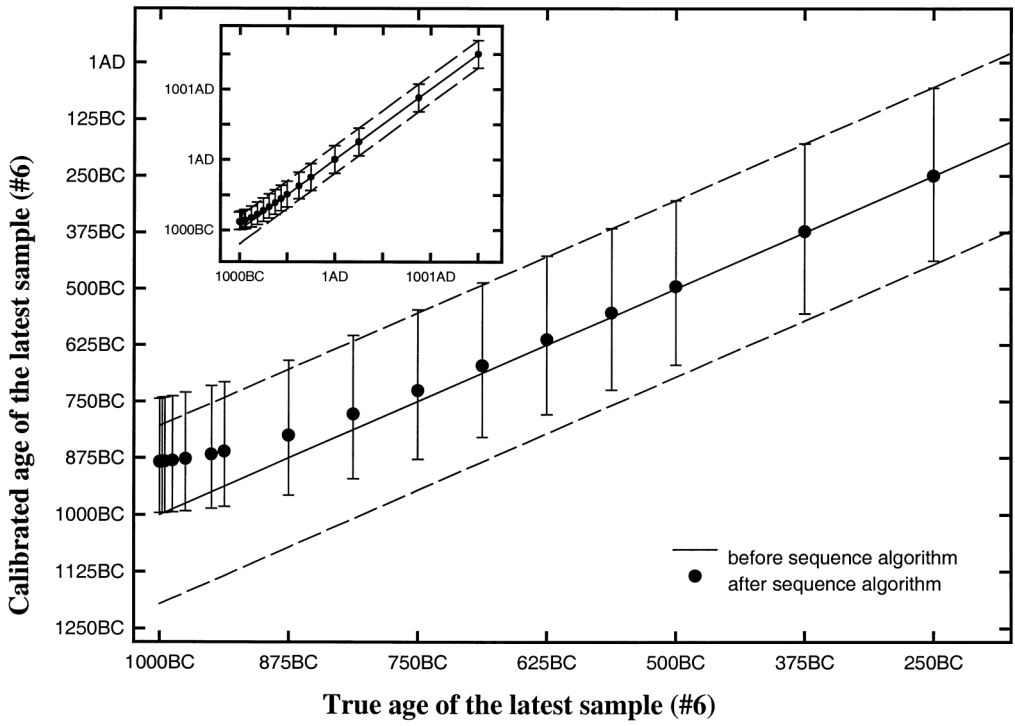

Figure 2 Computer experiment A. Maxima and 95% confidence intervals for the probability distributions of the latest sample #6 before (solid and dashed lines) and after the sequence algorithm (points and error bars) are shown. Only if $\Delta t \lesssim 2\sigma$ (i.e. t_6^{true} older than 750 BC) the data are changed significantly. The maximum of the posterior probability distribution is shifted away from t_6^{true} towards younger ages and the 95% confidence interval is incompatible with the assumed true age t_6^{true} = 1000 BC.

The random shifts ξ_k are obtained from a Gaussian probability distribution with a standard deviation equal to the assumed measurement uncertainty σ (100 yr). Around the shifted centers Gaussian-shaped probability distributions with the very same standard deviation of 100 yr were created (see Figure 3). Twenty such sets were generated and fed into the sequence algorithm. The resulting ages and uncertainties are shown in Figure 4. The strong shift of the age of the latest sample is further enlarged by the statistical scatter. This effect only accounts for a small part of the increased span.

To check whether realistic calibrated age distributions $P_k^{cal}(t_k)$, which are not Gaussian-shaped, influence the main features of the computer experiment we study a set of data typical for the "flat" case. For the whole Hallstatt period (750–400 BC, i.e. the Early Iron Age in Europe) the ^{14}C calibration curve is flat. Every sample yields a wide calendar age distribution with some wiggles, but the 95% confidence interval is very likely to span the whole period. We check the computer experiment for Δt = 5 yr. Figure 5 shows what happens when simulated ^{14}C ages of six Hallstatt samples (with typical measurement scatter) are used. The result is qualitatively the same as for the analytical investigation of Gaussian-shaped distributions. The centers of the distributions are shifted apart to cover the whole period, and the 95% confidence intervals are reduced so that the latest of the six samples is no longer compatible with the first half of the Hallstatt period.

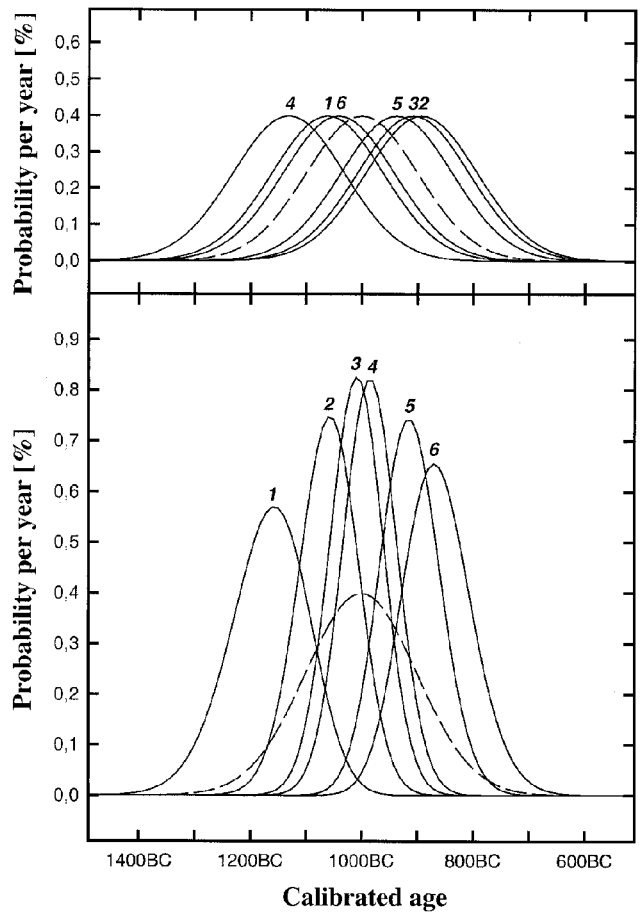

Figure 3 Computer experiment A. Modeling statistical scatter would not change the qualitative result shown in Figure 2. We check this for $\Delta t = 0$ yr by simulating sets with a randomly generated Gaussian scatter of $\sigma = 100$ yr. The probability distributions for the calibrated ages before (see upper part of the Figure) and after applying the sequence algorithm (see lower part of the Figure) are plotted.

In the Hallstatt period the posterior probability distributions will be essentially the same for any sequence independent of the (assumed) true ages. If the number of samples is sufficiently large, the latest sample is always shifted to 420 BC with a pretended small uncertainty.

Computer Experiment B

Next we study a growing number of samples N within a sequence. All samples are assumed to have the same true age ($\Delta t = 0$ yr) without measurement scatter. As can be seen in Figure 6 the sequence algorithm shifts the distributions more and more apart. The calibrated age range allowed by the ^{14}C measurement uncertainties is evenly partitioned between the posterior distributions. By increasing the number of samples the latest sample shifts to values deviating far from the assumed true ages (Figure 7).

Computer Experiment C

It can be seen from computer experiment B that the influence of the sequence algorithm grows with an increasing number of samples N in the sequence. For a small number of ordered samples there exist no obvious artifacts, but even in the case of two samples the uncertainties are significantly reduced. Is this increased precision accompanied by an increased accuracy?

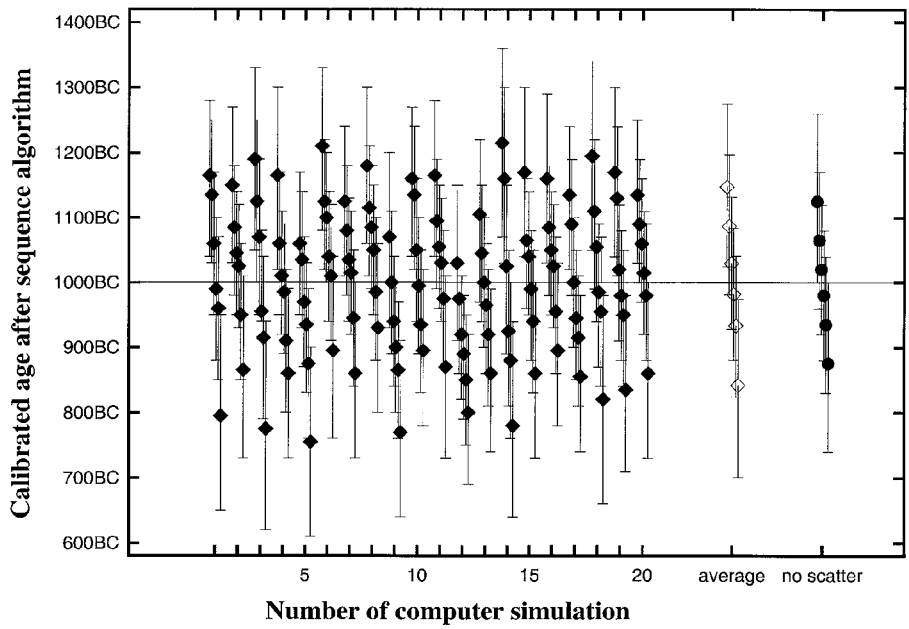

Figure 4 Computer experiment A. We compare the posterior centroids and 95% confidence intervals for 20 computer simulations including scatter constructed as in Figure 3 (filled diamonds) to the posterior data without scatter constructed as in Figure 1 (filled circles). The 95% confidence interval of sample #6 is not compatible with the assumed true age $t_6^{true} = 1000\,BC$ in 14 of the 20 cases. By averaging the centroids and the positive and negative interval widths we see an additional spread induced by the random scatter (hollow diamonds).

To answer this question we focus on ordered pairs of samples. When we repeat computer experiment A with just two samples in every set the influence is not as strong as the influence on multiple ordered samples (compare Figure 8 to Figure 2), but the ages are shifted apart also.

In computer experiment C we model statistical scatter. We simulate 1000 ordered pairs of samples for every assumed true age difference Δt. The true ages of the paired samples are symmetric around 1000 BC (see (10) with N = 2). The $P_k^{cal}(t_k)$ are modeled by using (12) with scatter and measurement uncertainty σ of 100 yr (see Figure 9). The artificial calibrated ^{14}C data together with the chronological ordering of the true ages is fed into the sequence algorithm.

If the algorithm really improves the dates, then the assumed true ages should be compatible with the posterior 95% confidence intervals in about 1900 of the 2000 cases (there are 2 samples for each of the 1000 pairs). The uncertainties shown in Figure 10 are induced by the binomial statistics of the experiment

$$u(m) = \sqrt{m \cdot \left(1 - \frac{m}{M}\right)} \tag{13}$$

where M is the total number of trials (2000) and m is the number of successful trials (number of 95% confidence intervals compatible with the assumed true age).

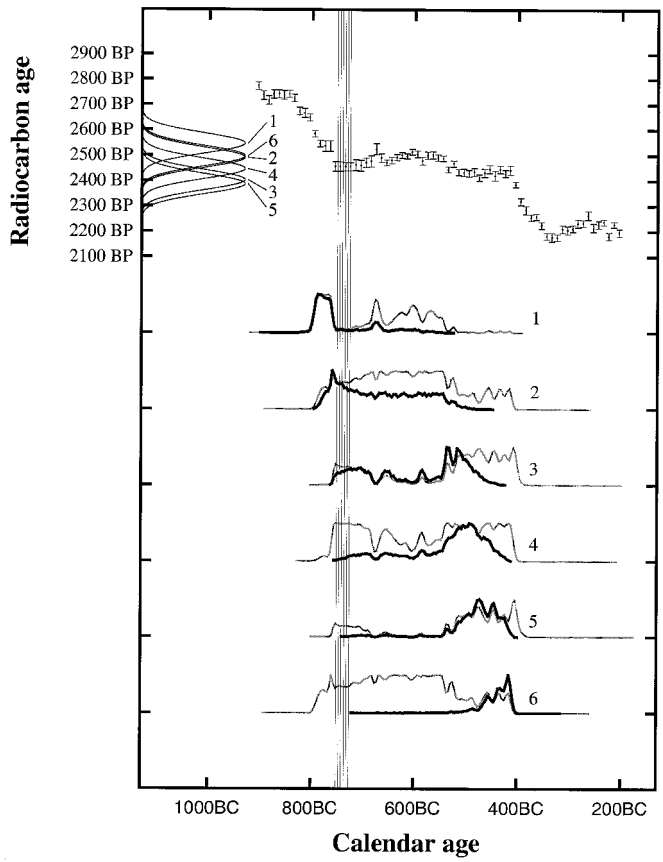

Figure 5 For an assumed set of 6 samples #1 to #6 from the Hallstatt period (750–400 BC) with ages of 750 BC (#1), 745 BC (#2), 740 BC (#3), 735 BC (#4), 730 BC (#5), and 725 BC (#6) indicated by vertical thin lines the corresponding ^{14}C ages were looked up in the calibration curve. Due to the flatness of the calibration curve we get the same ^{14}C age of 2455 BP for all 6 samples. After adding a random scatter of ±40 yr we obtain the following ^{14}C ages: 2546 BP, 2490 BP, 2402 BP, 2446 BP, 2386 BP, and 2491 BP. By individual calibration the samples can no more be assigned to distinct regions. The resulting probability distributions (gray curves) rather cover the whole Hallstatt period. These probability distributions correspond to our simulated ^{14}C measurement data. After the Bayesian sequence algorithm is applied one can see its tendency to divide the period into 6 parts of equal size (black curves). Due to the flatness of the calibration curve the general shape of the individually calibrated and of the "sequenced" probability distributions is the same which true ages ever are assumed. In our example the posterior 95% confidence intervals of samples #4, #5, and #6 are not in agreement with their assumed true ages. All the calculations (single calibration and sequencing) were performed with OxCal v2.18 (Ramsey 1995b) using the INTCAL98 ^{14}C calibration curve (Stuiver et al. 1998). The program normalizes the individual and the "sequenced" probability distributions to the same maximum value.

The number of simulated samples which miss their true ages before the sequence algorithm is applied shows no significant deviation from the theoretical 5% line. After applying the sequence algorithm the situation is changed drastically for true age differences smaller than the assumed measurement uncertainty. The posterior 95% confidence intervals miss the true ages in up to 12% of all cases, so the increased precision is an artifact. The fraction of incompatible intervals approaches the

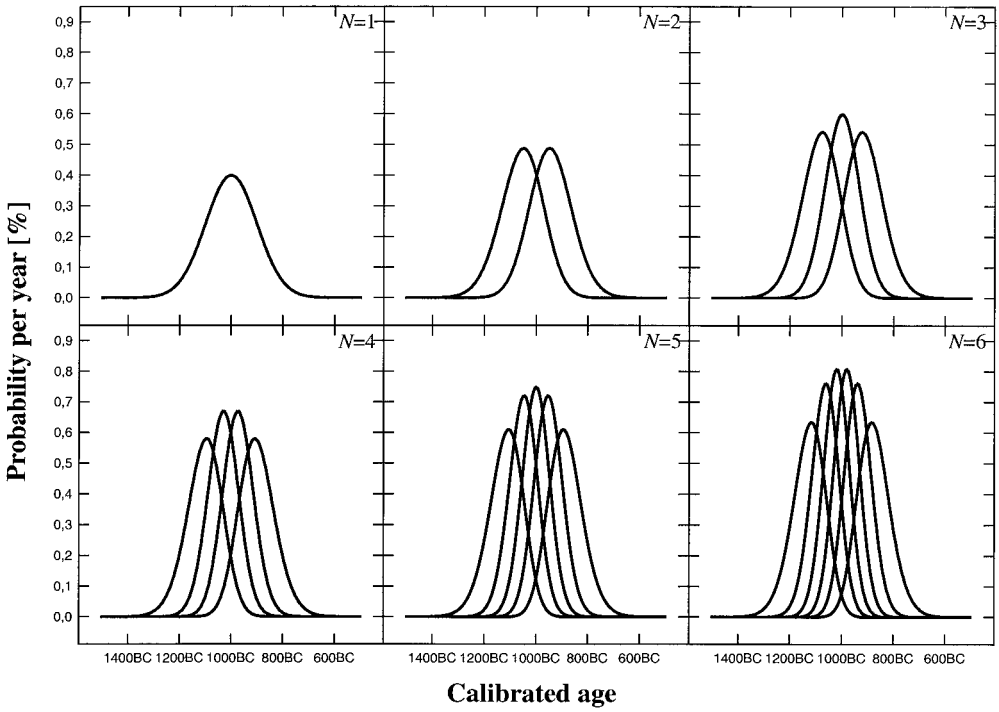

Figure 6 Computer experiment B. By increasing the number of samples N in a sequence they are more and more shifted apart by the Bayesian sequence algorithm. The individually calibrated probability distributions are all constructed Gaussian-shaped with centers at 1000 BC (Δt = 0 yr) and with σ = 100 yr. No scatter is modeled. The probability distributions after applying the sequence algorithm are plotted for N = 1 to N = 6 samples in a sequence.

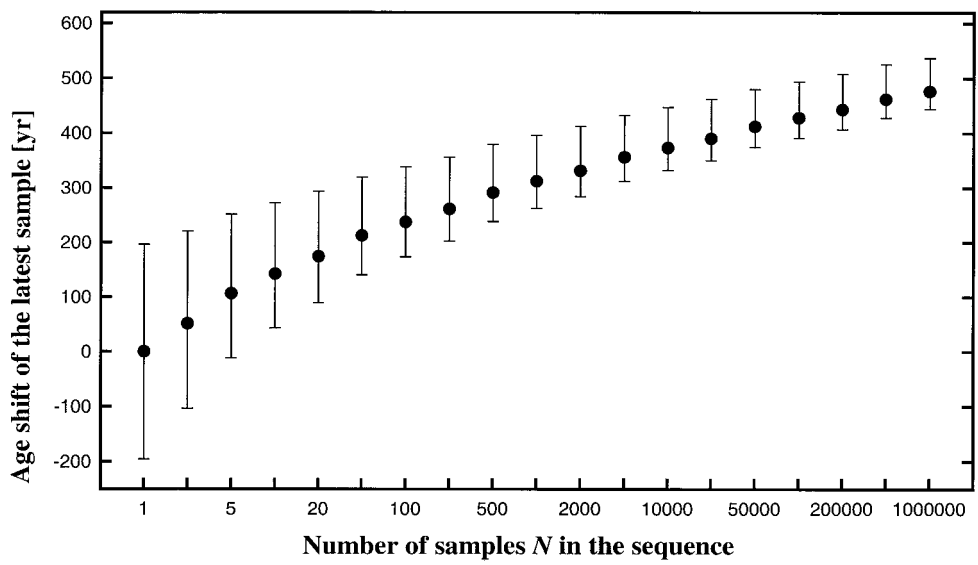

Figure 7 Computer experiment B. Maxima and 95% confidence intervals after applying the sequence algorithm are plotted for sets constructed as in Figure 6. The shift of the latest sample #6 grows with the number of samples N in a sequence (up to Ns too large to be realistic).

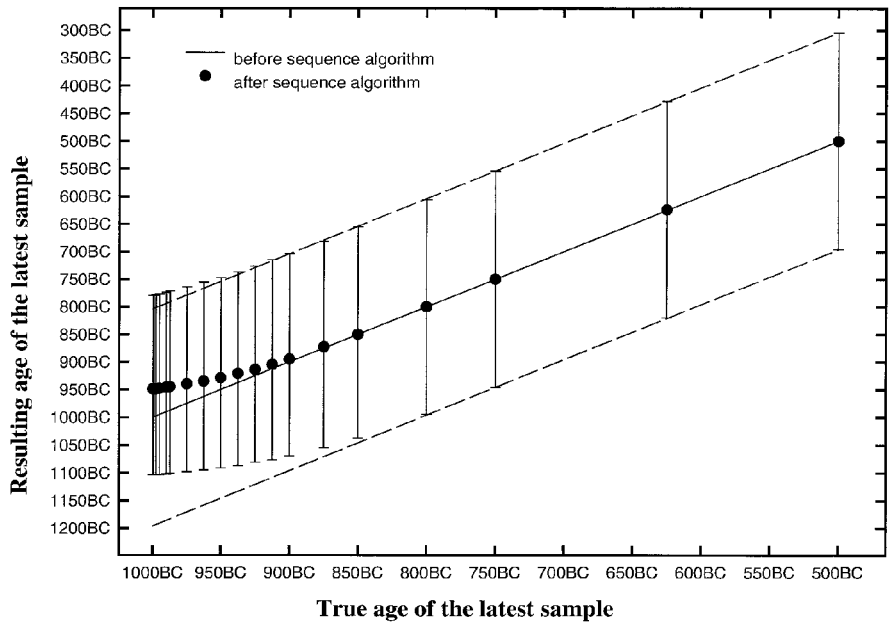

Figure 8 Computer experiment C. Whereas the artifacts in the results of the sequence algorithm are obvious for N = 6 or larger they are harder to detect for a smaller number of samples N in a sequence. The figure shows data analogous to Figure 2, but for N = 2.

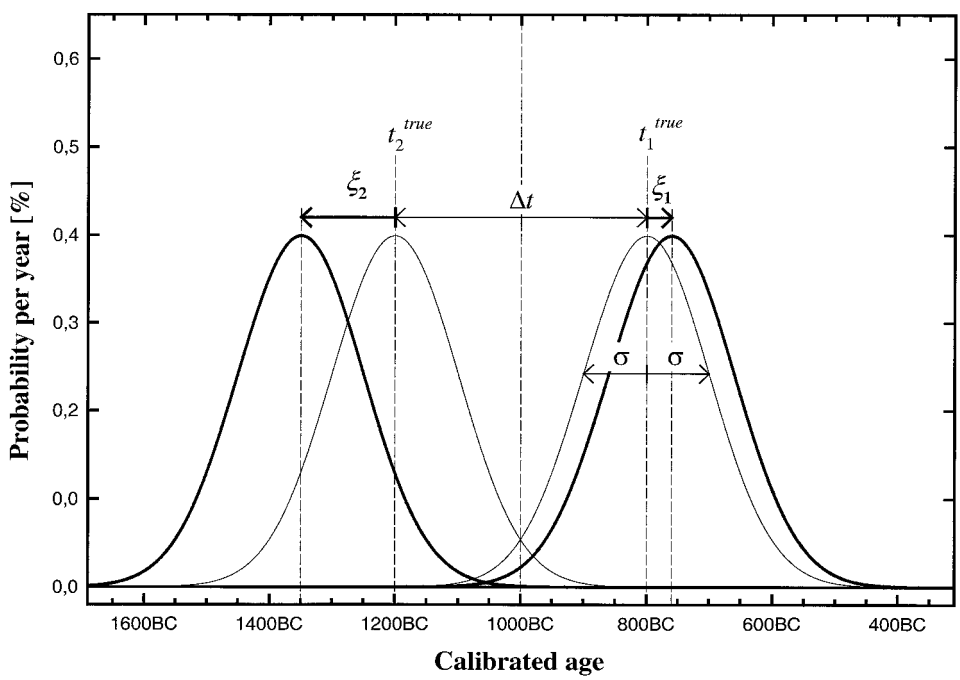

Figure 9 Computer experiment C. Pairs of samples are constructed symmetrically around 1000 BC with a time difference Δt between the true ages t_1^{true} and t_2^{true}. The individually calibrated probability distributions before applying the sequence algorithm are constructed Gaussian-shaped with σ = 100 yr. Measurement scatter is modeled by applying random shifts ξ_1 and ξ_2, which are taken from a Gaussian distribution also with σ = 100 yr.

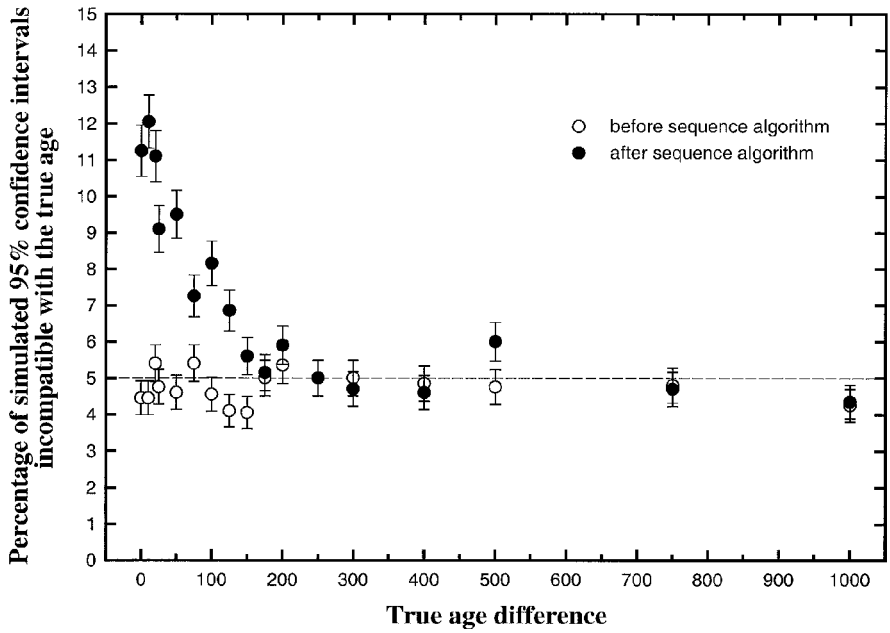

Figure 10 Computer experiment C. For each assumed true age difference Δt 1000 pairs of samples are constructed as in Figure 9, and the Bayesian sequence algorithm is applied. The 95% confidence intervals are checked for compatibility with the assumed true ages t_1^{true} and t_2^{true}. Whereas indeed only 5% of the single-sample calibration intervals are incompatible, after applying the sequence algorithm for $\Delta t < \sigma$ up to 12% of the intervals do not contain the corresponding true age.

5% line only if the difference Δt is larger than 1 σ (100 yr), but in this case the sequence algorithm has essentially no influence on the posterior probability distributions.

We want to emphasize that all the results in the computer experiments above scale with Δt/σ and in fact were calculated using only this parameter. For the figures shown they were scaled to values typical for ^{14}C dating.

'PRIOR' CONSIDERATIONS

We think the failure of the sequence algorithm as demonstrated by computer-simulated measurements is due to the prior probability assumed for the age difference of samples with known chronological order. Unfortunately the results of the sequence algorithm are highly sensitive to the assumed prior. The strong influence of the prior is demonstrated by choosing the probability as inversely proportional to the age difference (it is plausible that for samples from the same archaeological context smaller age differences are more probable). For two samples one gets:

$$P^{prior}(t_1, t_2) = \begin{cases} \dfrac{1}{t_2 - t_1} & \text{for } t_1 < t_2 - \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where ε < 1 is a small lower limit for the age difference to maintain integrability. This is equivalent to the assumption that there will be the same number of samples within 1 to 10, 10 to 100, and 100

to 1000 years of age difference. This principle of "scale invariance" (May 1996) holds for many positive numbers in nature. So this prior probability distribution may also be called "natural", but contrary to the constant probability the samples are drawn together instead of being shifted apart. For $\lim \varepsilon \to 0$ all ^{14}C data are shifted to the same age.

A constant probability may appear a neutral assumption, but every probability distribution gets constant by a suitable transformation of the variables. For example, (14) is constant if t_1 and t_2 are replaced by $\log t_1$ and $\log t_2$.

The sequence information restricts the age difference of consecutive samples to positive numbers and as pointed out by other authors, for positive numbers there is no reason to select a constant prior probability for the number itself and not for the logarithm or the square root of the number (see Blobel and Lohrmann 1998).

In the common prior (6) the ages t_k are selected as the "natural" parameters for which a constant probability is assumed. This is equivalent to the assumption that all dates—despite archaeologically related—are independent in a statistical sense. Another probably even more "natural" set of parameters would be the start time t_1 of the sequence and its total "span" $\Delta t_{1,N} = (t_N - t_1)$ together with the age differences $\Delta t_{1,k} = (t_k - t_1)$ of sample number k from the first. The common constant prior (6) is no longer constant for the $\Delta t_{1,k}$. By integrating over all combinations of $t_1, ..., t_{k-1}, t_{k+1}, ..., t_N$ and substituting $t_1 + \Delta t_{1,k}$ for t_k we obtain:

$$
\begin{aligned}
P^{\text{prior}}(\Delta t_{1,k}) &\propto \int_{-\infty}^{\infty} dt_1 \int_{-\infty}^{\infty} dt_2 \ldots \int_{-\infty}^{\infty} dt_{k-1} \int_{-\infty}^{\infty} dt_{k+1} \ldots \int_{-\infty}^{\infty} dt_N \cdot P^{\text{prior}}(t_1, \ldots, t_{k-1}, t_1 + \Delta t_{1,k}, t_{k+1}, \ldots, t_N) \\
&= \int_{-\infty}^{\infty} dt_1 \int_{t_1}^{t_1+\Delta t_{1,k}} dt_2 \ldots \int_{t_{k-2}}^{t_1+\Delta t_{1,k}} dt_{k-1} \int_{t_1+\Delta t_{1,k}}^{\infty} dt_{k+1} \ldots \int_{t_{N-1}}^{\infty} dt_N \cdot \text{const} \\
&\propto \int_{-\infty}^{\infty} dt_1 \int_{t_1}^{t_1+\Delta t_{1,k}} dt_2 \ldots \int_{t_{k-2}}^{t_1+\Delta t_{1,k}} dt_{k-1} \propto \int_{-\infty}^{\infty} dt_1 \; \Delta t_{1,k}^{\,k-2} \\
&\propto \Delta t_{1,k}^{\,k-2}
\end{aligned}
\qquad . \qquad (15)
$$

The common prior (6) considered "neutral" has a strong bias towards larger age differences if the number of samples exceeds 2. This problem strongly influences the calculation of the span $\Delta t_{1,N}$, i.e. the duration of the whole sequence. According to theorem (15) this has a marginal prior of $P^{\text{prior}}(\Delta t_{1,N}) \propto \Delta t_{1,N}^{N-2}$. The marginal priors of the samples show an analogous bias. This was already observed in Buck et al. (1991), but without discussing the implications, especially on the span calculation. Recently, Bronk Ramsey (1999) suggested to overcome this bias by modifying the common prior (6)

$$
P^{\text{prior}}(t_1, t_2, \ldots, t_N) = \begin{cases} \dfrac{1}{(t_N - t_1)^{N-2}} & \text{for } t_1, \ldots, t_N \text{ in order ("allowed case")} \\ 0 & \text{otherwise ("forbidden case")} \end{cases}
\qquad . \qquad (16)
$$

This can be achieved in the program OxCal by using the "BOUND" condition (C Bronk Ramsey, personal communication 1999). This condition is usually used to estimate the boundary (i.e. start and end time) of the sequence. In general, the significant gain in precision obtained with prior (6) cannot be achieved by using prior (16). A general use of the modified prior (16) may lead to an appreciable change of any posterior probability distribution calculated with prior (6).

We do not consider the modified prior (16) more generally valid than the common prior (6). Moreover, since for two samples the two priors (6) and (16) are identical, the failure in computer experiment C persists. Summing up we see no convincing way to select a certain prior for general use in the sequence algorithm.

There exist applications where the prior information is known in full detail, and therefore this information can be transformed into a mathematical form without vague assumptions. Here our criticisms concerning the sequence algorithm do not apply. This is the case for ^{14}C "wiggle matching" (Goslar and Wiesław 1998; Bronk Ramsey 1999) where the age differences for a set of samples, e.g. for N different tree rings from the same log, are known exactly. Each piece is ^{14}C-dated independently. The additional tree-ring information can be written as a prior probability distribution:

$$P^{prior}(t_1, t_2, ..., t_N) = \frac{1}{N-1}\left[\delta(t_2 - t_1 - \Delta t_{1,2}) \cdot \delta(t_3 - t_1 - \Delta t_{1,3}) \cdot ... \cdot \delta(t_N - t_1 - \Delta t_{1,N})\right] \quad (17)$$

where time offset $\Delta t_{1,k}$ can be obtained from the number of tree rings in between. No vague assumptions like the constant probability in the common prior (6) have to be made.

CONCLUSION

Bayesian mathematics is a powerful tool for combining probability distributions from different sources, if these distributions are well defined. In this paper we discussed the combination of a well-defined distribution derived from ^{14}C measurements of archaeological samples with additional information on their chronological order. This information can only be transformed into complete probability distributions by using "vague" assumptions. The prior commonly used is a constant probability density for the calendar ages, as long as the given order is respected (otherwise it is zero).

If the samples are already well resolved in time by the ^{14}C measurement alone, the Bayesian sequence algorithm does not change the data. For samples that cannot be separated by the ^{14}C measurement we demonstrated by means of computer-simulated measurements that the common prior creates results with spurious high precision. The algorithm spreads the ages of the samples in a sequence over the whole range allowed by the ^{14}C uncertainty (which may be large for flat regions of the calibration curve), and small uncertainties are obtained. These results are no longer in agreement with the (assumed) true ages of the samples and therefore the reduced uncertainties are an artifact of the algorithm. Generally speaking, the algorithm improves the precision but reduces the accuracy! We demonstrated that these problems show up in any region of the calibration curve. The artifacts are more obvious for a larger sequence of samples but even persist for only two samples.

We came to the conclusion that the commonly used prior is no "neutral" assumption. The decision which prior probability distribution is suited for the individual archaeological context should be made in close cooperation with archaeologists well-experienced in quantitative methods.

ACKNOWLEDGMENT

We would like to thank Christopher Bronk Ramsey from the Oxford Radiocarbon Accelerator Unit for initiating our interest in the subject of this paper through his talk "From radiocarbon measurement to chronological interpretation" at our weekly laboratory seminar in December 1997, and for various discussions continued via e-mail and fax, and personally at the 8th International Conference on Accelerator Mass Spectrometry in Vienna, September 1999.

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.

Bayliss A, Bronk Ramsey C, McCormac FG. 1997. Dating Stonehenge. In: Cunliffe B, Renfrew C, editors. *Science and Stonehenge*. Oxford: Oxford University Press. The corresponding OxCal program code is accessible at URL: <http://www.eng-h.gov.uk/stoneh/codemain.htm>.

Blobel V, Lohrmann E. 1998. *Statistische und numerische Methoden der Datenanalyse*. Stuttgart; Leipzig: B.G. Teubner. p 204–9.

Bronk Ramsey C. 1995a. Radiocarbon calibration and analysis of stratigraphy: the OxCal program. *Radiocarbon* 37(2):425–30.

Bronk Ramsey C. 1995b. OxCal Program v2.18. URL: <http://units.ox.ac.uk/departments/rlaha/oxcal/oxcal_h.html>.

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

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.

Buck CE, Litton CD, Scott EM. 1994. Making the most of radiocarbon dating: some statistical considerations. *Antiquity* 68:252–63.

Buck CE, Litton CD, Smith AFM. 1992. Calibration of radiocarbon results pertaining to related archaeological events. *Journal of Archaeological Science* 19:497–512.

Buck CE, Kenworthy JB, Litton CD, Smith AFM. 1991. Combining archaeological and radiocarbon information: a Bayesian approach to calibration. *Antiquity* 65: 808–21.

Goslar T, Wiesław M. 1998. Using the Bayesian method to study the precision of dating by wiggle-matching. *Radiocarbon* 40(1):551–60.

Litton CD, Buck CE. 1995. Review article – The Bayesian approach to the interpretation of archaeological data. *Archaeometry* 37(1):1–24.

May RM. 1996. Wie viele Arten von Lebewesen gibt es? In: König B, Linsenmair KE, editors. *Biologische Vielfalt*. Heidelberg, Berlin, Oxford: Spektrum Akademischer Verlag GmbH. p 16–23.

Reece R. 1994. Are Bayesian statistics useful to archaeological reasoning? *Antiquity* 68:848–50.

Roe BP. 1992. *Probability and statistics in experimental physics*. New York: Springer-Verlag. p 103–5.

Stuiver M, Reimer PJ, Bard E, Beck JW, Burr GS, Hughen KA, 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.