16

Some statistical problems arising in radiocarbon calibration

C. D. Litton

(Department of Mathematics, University of Nottingham, Nottingham, NG7 2RD)

M. N. Leese

(Research Laboratory, The British Museum, London, WC1B 3DG)

16.1 Introduction

This paper is concerned with the statistical analysis of sets of results in radiocarbon dating, specifically where the events dated have some *a priori* relationship between them, either known or assumed. Two statistical models for the single-date situation are contrasted and one of these is shown to be readily extended to the analysis of more complex situations. Firstly, however, we discuss the reasons why an examination of this topic is timely and outline the general strategy with which it we feel it should the approached.

The statistical treatment of sets of calibrated radiocarbon dates has recently become an issue of importance with radiocarbon laboratories increasingly encouraging the submission of groups of samples relating to an archaeological context rather than just single samples. As such the radiocarbon results require integrating with the archaeological evidence. There seems to be no accepted way of dealing with such collections of radiocarbon results, other than by various *ad hoc* methods, and even calibrating a single result appears to cause confusion. That there is no consensus on exactly how the calibration should be performed was evident from a recent survey of radiocarbon calibration programs (Aitchison *et al* 1989); different programs produced different results for the same radiocarbon result; some programs could only cope with multiple dates on the same sample, while others combined dates for different events. Some variations among the programs were due to different approaches to smoothing the curve, incorporating the error in the curve itself, and other important but secondary considerations, but it seems that differences of opinion regarding methods stem mainly from the use of different statistical models. These models are not generally explicitly stated and this, we believe, is why there is great confusion about the correct approach, quite apart from the mathematical problems caused by the complex form of the calibration curve.

The mathematical difficulties occur because the high precision calibration curve which is now internationally accepted (Stuiver & Pearson 1986, Pearson & Stuiver 1986) is non-monotonic or *multiple intercept*, as shown in Fig. 16.1. As noted above there are a number of computer programs for dealing with the complexities of the curve, but somewhat surprisingly there has been little pressure from archaeologists to develop a more unified and rational approach, given that the programs are all slightly different, especially in their treatment of sets of dates. It may be that there is a general feeling that the errors in the radiocarbon dates are too large, and the number of dates too small, to justify sophisticated approaches to their interpretation. Such a view was expressed at a recent one-day conference at the British Museum. Nevertheless, we will argue that it is *only* by adopting a correct statistical model which incorporates all the available information into the analysis that valid interpretation can be made.

If this principle is accepted then, for a particular situation under consideration, attention must be paid to the following points.

1. The appropriate statistical model must be formulated. This must be carried out in collaboration with the archaeologist concerned.
2. The correct assessment of all the sources of random variation in the radiocarbon result must be made in collaboration with radiocarbon experts.
3. Suitable statistical procedures will be needed to estimate the parameters of the model(s) developed in 1. These will have to be embodied within user-friendly computer software.
4. The archaeological community initially will need help and guidance with the interpretation of the results.

Error estimation is mainly the province of radiocarbon laboratories, and while a re-examination of some of the assumptions about the forms of the errors may be advisable, in this paper we concentrate on the other three points. It will be obvious from this that we differ from the view of Aitchison *et al.* (1989) who adopt the *floruit*, or period over which 50% of a culture's datable artefacts were produced, as the single quantity of interest. On the contrary we regard it as necessary to formulate the model afresh for each situation, under guidance from the archaeologists.

One point which should be made at this stage is that while measurement bias in the radiocarbon result must of course be identified if possible and eliminated, an age offset due to the use of mature wood (also regarded as a kind of bias) affects the true calendar date and not the radiocarbon result. In principle, therefore, it should be regarded as a possible input to the model and not as another source of 'error'. Other types of bias in the archaeological record and its sampling may not be so readily identified and hence incorporated into the analysis; the possibility of these must always be borne in mind. Also it is worth distinguishing between multiple dates which relate to the same event and those which relate to different events. It is the latter with which we are concerned: replicate dates on the same sample or dates from separate samples from the same object, or in some cases, the same archaeological context, can be examined and combined at the radiocarbon level before calibration since they have the same true radiocarbon result, and calibration to the calendar scale is irrelevant at this stage (see e.g. Bowman 1990).

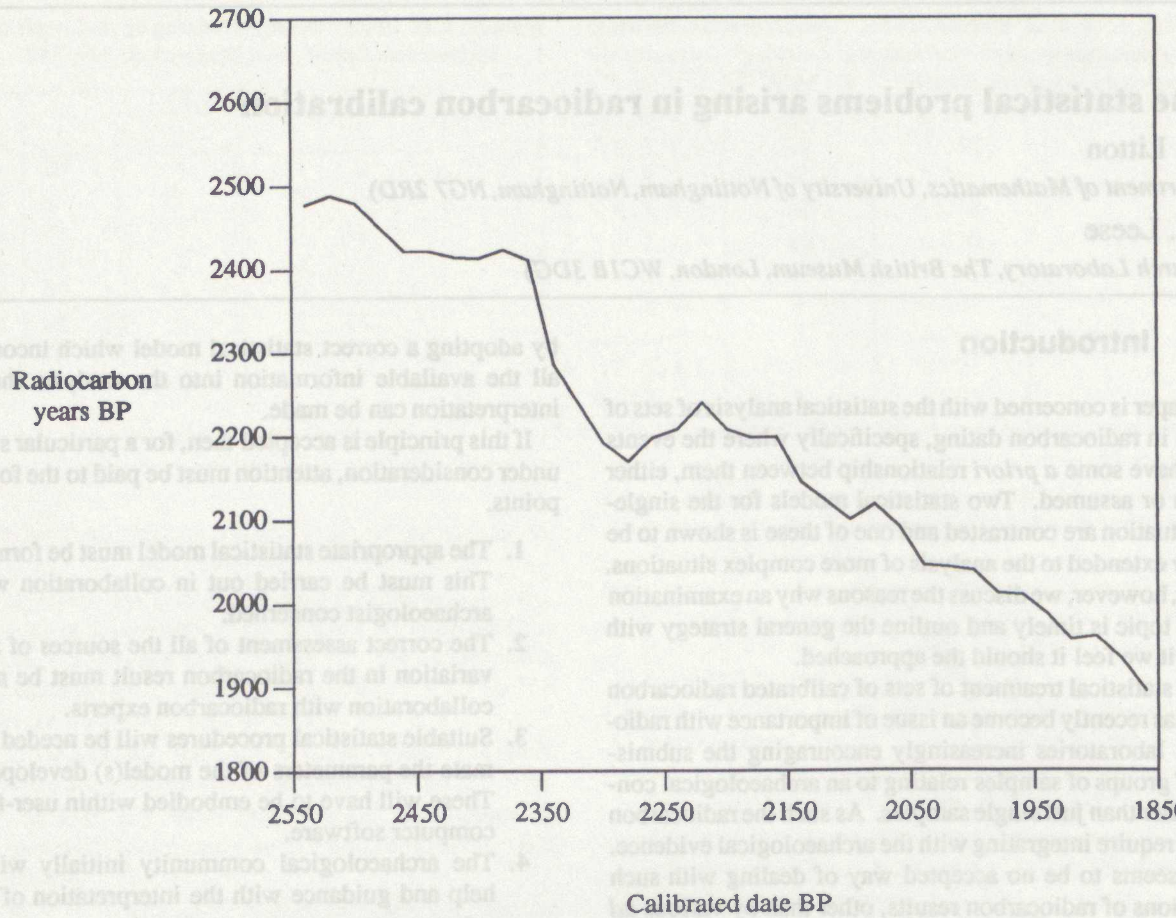

Figure 16.1: Part of the high precision calibration curve (Pearson & Stuiver 1986, Stuiver & Pearson 1986) showing typical wiggles. Error bands on the curve have been omitted for clarity.

16.2 Previous treatment of multiple dates.

A few publications deal with statistical models for dates relating to different events, and a selection of these illustrate the range of situations which can arise. There is clearly a need for a general approach which is flexible enough to be adapted to a wide variety of situations.

One of the earliest attempts to analyse a complex set of dates was that of Orton (1983). He developed a model for a number of possibly overlapping phases, from each of which a number of radiocarbon results were available, placed in phases on the basis of ceramic associations; the aim was to estimate the beginnings and endings of the phases on the radiocarbon scale. This was later re-examined by Naylor and Smith (1988) from a Bayesian viewpoint which we will discuss in depth later. Helskog and Schweder (1989) considered a number of Late Stone Age Norwegian houses, each having an associated radiocarbon result. They estimated, on the basis of the dates and their errors, the probability distribution that each house was occupied at a given time, and hence the number of houses simultaneously occupied. A Bayesian approach to the case where a set of dates is available for each house was also suggested. Vincent (1989) also suggested a Bayesian approach, in this case for dealing with the common situation in which one date is known to be older than a second on the basis of stratigraphy although the radiocarbon results are in the reverse order. Apparent contradictions of this kind could arise quite naturally with multiple intercept curves as shown by Bowman (1990). One can envisage many more possibilities, for example models including constraints introduced by *termini ante quem* and *termini post quem* implied by stratigraphic or historical evidence.

Clearly such studies are relevant to real archaeological problems but their restriction to monotonic calibration curves or radiocarbon results before calibration limits their applicability. A study which combined the use of several multiple intercept curves (but not the now accepted version) with a realistic problem (the estimation of phase lengths given ceramic associations) was described by Naylor and Smith (1988). Their basic model together with the Bayesian methodology for parameter estimation provides a starting point for the study of other situations. Before we describe their approach we briefly discuss the model currently adopted, albeit implicitly, by radiocarbon laboratories in interpreting radiocarbon dates. There is a subtle difference between this model, which essentially describes the measurement process alone, and one which captures the whole process of the generation of the calendar event and its realisation and measurement in radiocarbon terms. It is this difference which we believe is the source of confusion in calibrating single dates and the reason why multiple date situations have not always been realistically treated.

16.3 The laboratory interpretation of radiocarbon error distributions.

Radiocarbon results are generally reported as $x \pm \sigma$ where σ is the standard deviation of the error. The evaluation of the error is partly theoretical and partly empirical; despite its empirical component, it is treated as if it were known, and confidence limits for the true radiocarbon date, X, are set up, based upon the implicit model

$$X = x + \mathrm{e}, \text{ where } \mathrm{e} \sim N(0, \sigma^2).$$

This is the model adopted by Ward and Wilson (1978) whose tests for homogeneity of results and recommendations for combining replicate samples are widely used amongst the radiocarbon community. From the point of view of the laboratory issuing the results, this model and interpretation are quite adequate since at the laboratory stage only the results of the measurement process are being expressed.

With this model, the measured radiocarbon result, x, is used as the best estimate of the mean of the normal distribution of possible radiocarbon measurements which might arise on repeated measurements of the same sample, i.e.

$$X \sim N(x, \sigma^2).$$

This model is apparently adopted by most programs for 'probability' calibration, including Robinson (1978), Leese (1988), and the CALIB program of Stuiver and Reimer (1986,1989), though as noted below the use of a 'non-dividing' probability method in the latter suggests the implicit use of a different model. If this model is adopted as it stands, it is necessary to divide the probability density among the various possibilities, caused by multiple intercepts, when transferring from the radiocarbon scale to the calendar scale. This can give rise to computational problems, and moreover this way of looking at the situation is not amenable to the extension to multiple dates since it has no way of incorporating any prior knowledge about the calendar dates sampled. A different approach is required for this. While the computations involved in multiple date extensions of the model below are difficult because of the requirement of numerical integration, the basic models are quite simple to formulate. However they do require a different approach to the model building process which is now explained.

16.4 The underlying statistical model for radiocarbon dated events.

The commonly held view of the conversion from a radiocarbon result to a calendar date is, given a distribution on the radiocarbon scale, usually expressed in terms of a radiocarbon date and its standard deviation, what is the corresponding distribution on the calendar scale? We believe that this emphasis on converting a *distribution* from one scale to another is misleading and that it is beneficial to view the problem as one of *estimating* the calendar date corresponding to an observation on the radiocarbon scale by means of a likelihood function. One of the consequences of changing this emphasis is that the resulting model readily lends itself to extension to the more complex multi-sample or multi-phase situations that were mentioned earlier.

We now formulate our model by assuming that we are attempting to date an event which has a *unique* but unknown calendar date, denoted by θ. Associated with this event is a unique radiocarbon date, $\mu(\theta)$; we are unable to measure $\mu(\theta)$ with total accuracy but we can observe a noisy version of this radiocarbon date, denoted by x. We can view this x as a realisation of a random variable X where

$$X = \mu(\theta) + \text{noise}.$$

Assuming that the noise has a normal distribution with mean zero and standard deviation σ then $X \sim N(\mu(\theta), \sigma^2)$. With this formulation it is clear that when a radiocarbon result is reported as $x \pm \sigma$, then this should be interpreted as a 68% confidence interval for the true radiocarbon date $\mu(\theta)$. Note that this is virtually the same as the laboratory interpretation, except for the now explicit dependence on θ.

We now have to relate $\mu(\theta)$ to θ using the high precision calibration curve expressed in its piecewise linear form

$$\mu(\theta) = \begin{cases} a_1 + b_1\theta & (\theta < \theta_0) \\ a_k + b_k\theta & (\theta_{k-1} < \theta < \theta_k, k = 1, 2, \ldots, K) \\ a_K + b_K\theta & (\theta > \theta_K) \end{cases}$$

where $K + 1$ is the number of knots used to define the calibration curve and where a_i and b_i are assumed to be known constants. Then we have

$$\begin{array}{ll} X \sim N(a_1 + b_1\theta, \sigma^2) & (\theta < \theta_0) \\ X \sim N(a_k + b_k\theta, \sigma^2) & (\theta_{k-1} < \theta < \theta_k, k = 1, 2, \ldots, K) \\ X \sim N(a_K + b_K\theta, \sigma^2) & (\theta > \theta_K) \end{array}$$

We are now in a position to write down the likelihood function relating the observed radiocarbon result, x, to the unknown calendar date, θ. It is simply

$$l(x;\theta,\sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma}\exp - \frac{(x-a_1-b_1\theta)^2}{2\sigma^2} & (\theta < \theta_0) \\ \frac{1}{\sqrt{2\pi}\sigma}\exp - \frac{(x-a_k-b_k\theta)^2}{2\sigma^2} & (\theta_{k-1} < \theta < \theta_k, k = 1, 2, \ldots, K) \\ \frac{1}{\sqrt{2\pi}\sigma}\exp - \frac{(x-a_k-b_K\theta)^2}{2\sigma^2} & (\theta_K < \theta) \end{cases}$$

To illustrate the nature of the likelihood we consider a highly simplified calibration curve with one wiggle. Plots of the likelihood for three different radiocarbon results (all with standard deviation of 100) are given in Fig. 16.2. For the result well away from the wiggle the likelihood is uni-modal and symmetric. However this is not so for results near the wiggle where the likelihood becomes decidedly asymmetric and bi- or even tri-modal. If we use the actual high precision calibration curve then the situation will obviously become more complex.

Before being able to make inferences concerning the unknown date, θ, we have to decide which statistical methodology to adopt. In the classical approach we would use the method of maximum likelihood which chooses as our best estimate of θ that value which results in the likelihood having a maximum. However it is clear that this is highly unsuitable for the multimodal likelihoods that arise in this problem. Alternatively we may use the Bayesian approach which can deal with the complexities arising from such likelihoods and also provides a means of including prior information in the analysis. To make inferences regarding θ we need to make prior assumptions about θ and this is expressed in terms of the prior probability density $p(\theta)$. Then the posterior probability density of θ which encapsulates both the prior information and the information carried by the radiocarbon results, is given by

$$p(\theta|x) = \frac{l(x;\theta)p(\theta)}{\int l(x;\theta)p(\theta)\mathrm{d}\theta}$$

Little or no prior knowledege about the calendar date may be expressed by a 'vague prior', i.e., by letting $p(\theta) = 1$ over the whole range of θ. Effectively this is saying that the event under consideration is, in the absence of any other information, equally likely to have occurred in any year. (This is, in fact, what the CALIB program assumes although it is not entirely clear that users of the program appreciate that this assumption has been made).

Of course in some situations more prior information will be available. For example we may know that a date should be, based on other evidence, between two calendar dates θ_a and $\theta_b(\theta_a < \theta_b)$ but that there is no evidence suggesting that one date within this range is more likely than any other. This may be captured by setting our prior density as

$$p(\theta) = \begin{cases} (\theta_b - \theta_a)^{-1}, & (\theta_a < \theta < \theta_b), \\ 0 & (\text{otherwise}) \end{cases}$$

Another common problem is that we know *a priori* that one event must be later than another and we have radiocarbon results for both. Let θ_1 and $\theta_2(\theta_1 < \theta_2)$ be the unknown calendar dates of the two events. Let x_1 and x_2 respectively be the corresponding observed radiocarbon results. Furthermore let the reported standard deviations be σ_1 and σ_2 respectively. Assuming that the radiocarbon results are independent of each other, the likelihood relating the observations x_1 and x_2 to the unknown dates θ_1 and θ_2 is given by

$$l(x_1, x_2; \theta_1, \theta_2, \sigma_1, \sigma_2) = l(x_1; \theta_1, \sigma_1)l(x_2; \theta_2, \sigma_2)$$

To make inferences regarding θ_1 and θ_2 we need to examine their joint posterior density.

Vincent (1988) has suggested a Bayesian approach to this problem and here we reanalyse an example he gave. For direct comparisions to be made with his results we assume that the calibration curve in the timespan of interest contains no wiggles and so is a straight line. The example has the radiocarbon results apparently in a different order to the known stratigraphic order.

Radiocarbon result	x_1	1900BP±85
Radiocarbon result	x_2	1800BP±70

Let the calendar dates of events 1 and 2 be θ_1 and $\theta_2 = \theta_1 + \theta_3$ respectively where $\theta_1 > 0$ and $\theta_3 > 0$. This formulation ensures that the event 2 is earlier on the calendar scale than event 1 despite the radiocarbon results suggesting otherwise. Assuming vague prior information over the region $\theta_1 > 0$ and $\theta_3 > 0$, plots of the joint posterior density of θ_1 and θ_3, their marginal densities and the marginal density of $\theta_2 = \theta_1 + \theta_3$ are given in Fig. 16.3.

The marginal posterior density of θ_1 is symmetric with a mode at 1810; its expected value and standard deviation are 1807 and 62 respectively. The marginal posterior density of θ_2 has a mode at about 1860; its expected value and standard deviation are 1863 and 58 respectively. These results show good agreement with those of Vincent obtained by a slightly different method. If we follow the convention of reporting two-sigma intervals these are 1683 to 1931, and 1747 to 1979 for θ_1 and θ_2 respectively. But these intervals overlap which contradicts our prior knowledge that θ_1 is smaller than θ_2. Thus it is obvious that reporting in this way will be grossly misleading and further analysis is required.

The contour plot of the joint posterior density of θ_1 and θ_3 is more informative, the mode is in the region of $\theta_1 = 1830$ and θ_3 very close to zero. That is the most likely dates for events 1 and 2 are about 1830BP and a little before this, respectively. If we want to report an interval for θ_1 and θ_3, and we encourage this, we need a means of summarising the shape of the contours. This is a somewhat difficult task and we believe that a summary in the form of a contour plot together with some brief explanation is probably adequate. Turning to the marginal plot of θ_3, this confirms that the

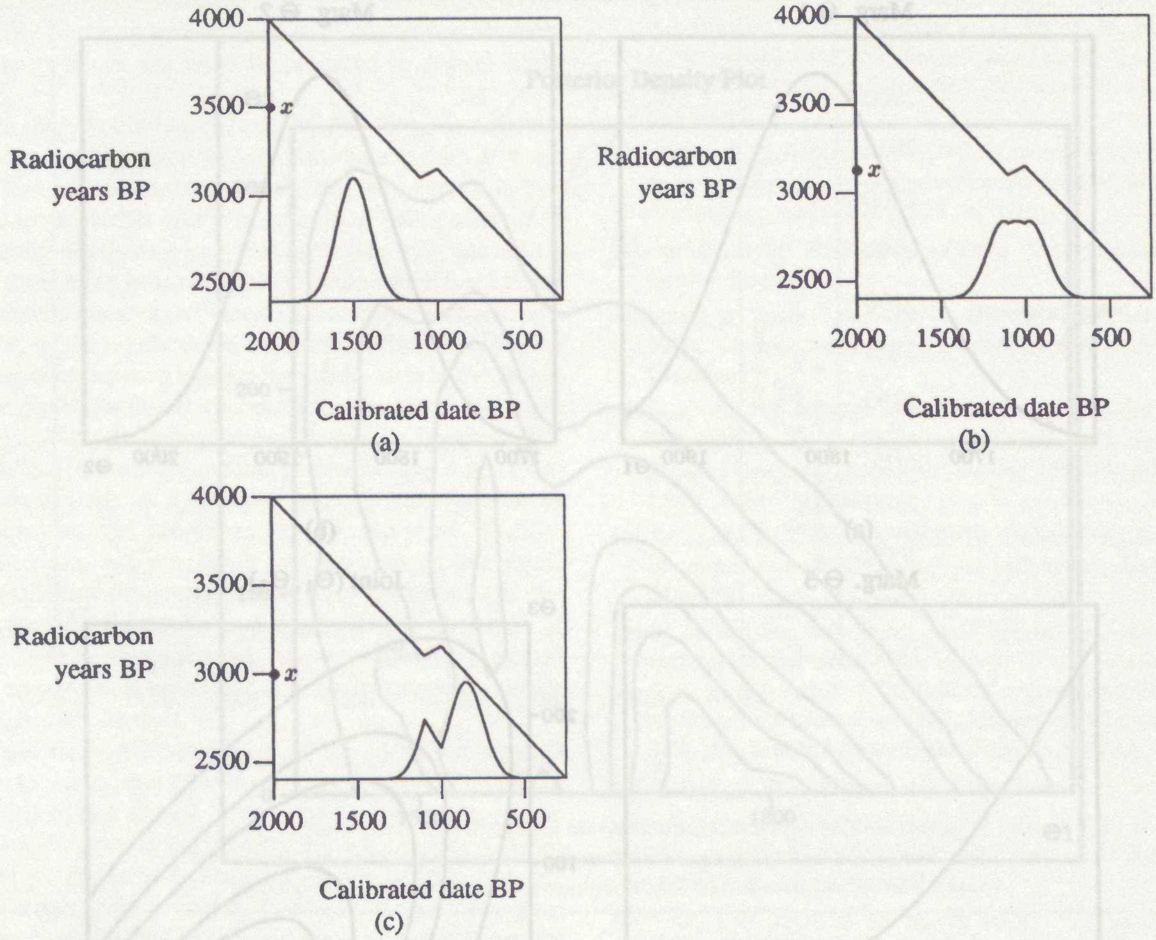

Figure 16.2: Plots of the likelihood using a simplified calibration curve

(a) $x = 3500, \sigma = 100$
(b) $x = 3125, \sigma = 100$,
(c) $x = 3000, \sigma = 100$

second event is most likely to have occurred just before the first, athough it could be up to 200 years earlier (depending on the date of the first event).

The above analysis illustrates the need to consider the joint posterior density for this type of situation. Looking at the two-way plot shows that inferences made about the second date (i.e., $\theta_2 = \theta_1 + \theta_3$) depend upon θ_1. At $\theta_1 = 1900$BP the fall off is very steep, whereas at 1700BP a very wide range of values for the second date are almost equally likely; in other words the marginal plots are not telling the whole story: we need the contour plot of the joint density to see how the two dates are related.

As we have just demonstrated the interpretation of results of this nature may can difficult even when the calibration curve used has no wiggles. Just for comparison purposes, plots of the joint posterior density of θ_1 and θ_3 using the high precision curve is given in Fig. 16.4. We leave the interpretation to the reader! While the interpretation of plots such as these may be initially difficult (and as we have suggested might require some statistical advice) they do give a realistic idea of the range of possibilities and their interaction. Moreover, if specific questions are posed then this kind of model can provide specific answers, even for quite complex situations.

16.5 Extension to more complex problems.

Turning to more complicated situations, we believe that the way forward has been pioneered by Naylor and Smith (1988). Ironically they were not interested in problems associated with radiocarbon dates *per se*, but wanted a complex problem, with a high dimensional, multimodal likelihood with constraints on the parameter values, with which to demonstrate the capabilities of their recently developed computer package, BAYES4 (Naylor & Shaw 1985). Their paper is written for statisticians and most archaeologists (and some statisticians) will not be able to appreciate its significance. For this reason we attempt to summarise the features that have important implications for archaeologists. To help do so we consider the paper in five different aspects, which will be common to all types of problem, namely:

1. the archaeological problem
2. the statistical model
3. the prior information
4. the statistical inference procedures
5. the interpretation of the results.

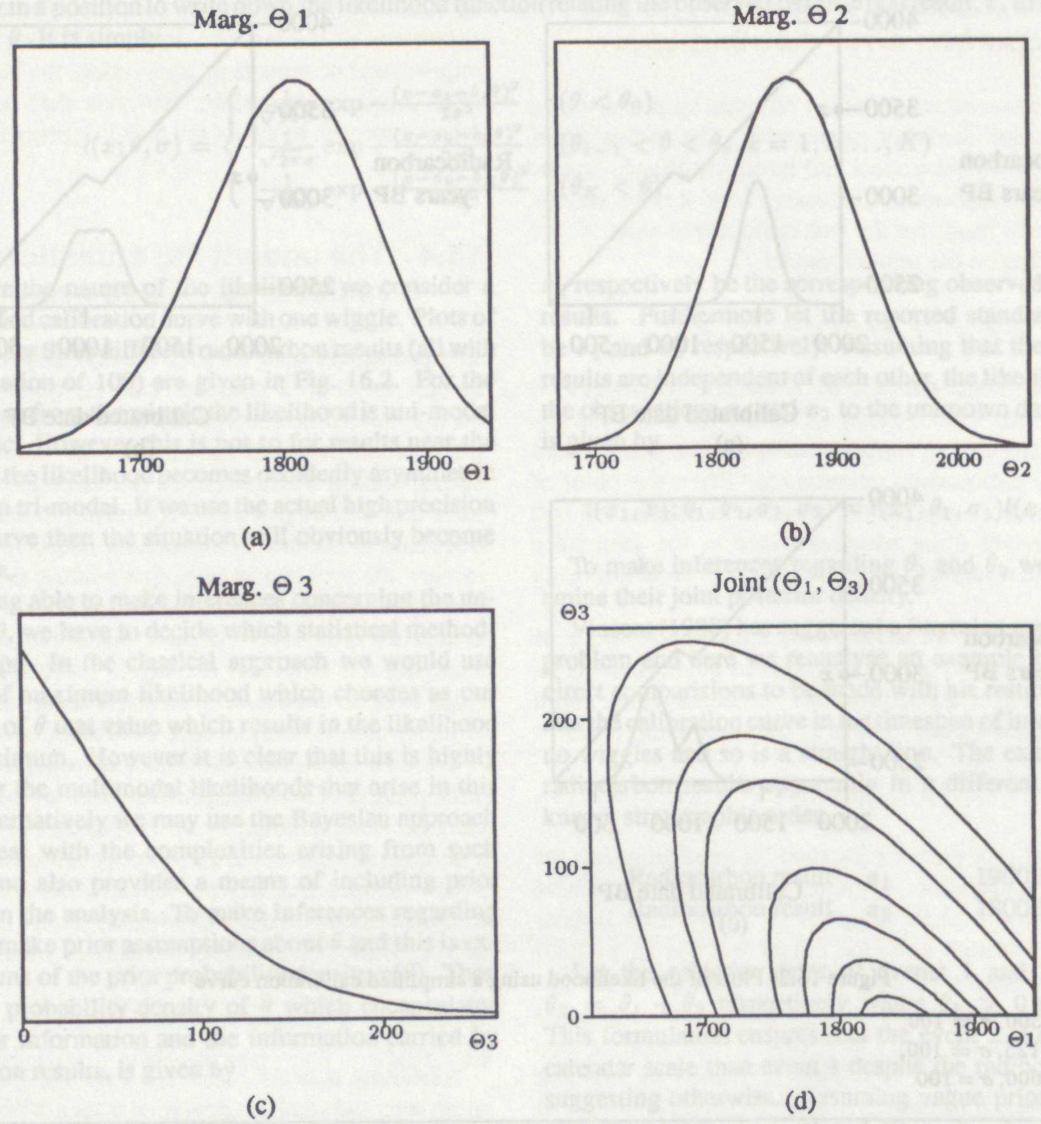

Figure 16.3: Plots obtained using a simplified calibration curve

(a) Marginal posterior density of θ_1.
(b) Marginal posterior density of θ_2.
(c) Marginal posterior density of θ_3.
(d) Joint posterior density of θ_1 and θ_3.

1 The archaeological problem

The problem relates to a collection of sixty-five radiocarbon results (with their associated measurement standard deviations) from the excavation of the Danebury iron-age hill fort (see Cunliffe 1984). Associated with the archaeological contexts from which the radiocarbon samples were taken are assemblages of archaeological finds. These finds include pottery sherds which ceramic experts have characterised into four phases. For convenience we will label these Phases 1–4. The problem is to determine the beginning and ending dates, on the calendar scale, of the four phases using the radiocarbon results. The radiocarbon results associated with each Phase and the relevant section of the high precision calibration curve which was used are given in Fig. 16.5.

2 The statistical model

In order to develop a statistical model appropriate to the question posed above, several modelling suppositions will have to made. The validity of these will obviously be questioned by archaeologists but we leave our discussion of this point until later.

1. For a sherd of calendar date θ the corresponding radiocarbon result is assumed to be a realisation from a normal distribution with mean $\mu(\theta)$ and known standard deviation σ.
2. The events corresponding to the radiocarbon results are statistically independent of each other. Also the radiocarbon determinations are statistically independent of each other.
3. The parameters of the high precision curve are known exactly.

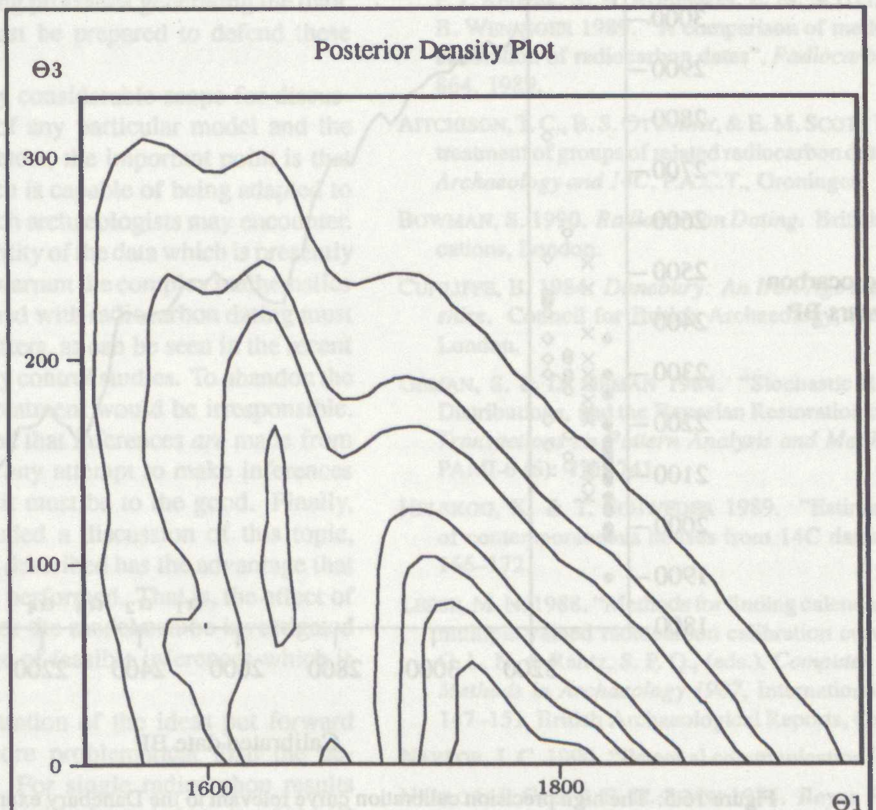

Figure 16.4: Joint posterior density of θ_1 and θ_3 using the high precision calibration curve

4. The four phases are regarded as abutting non-overlapping phases. Let α_1 be the beginning of Phase 1, α_2 the ending of Phase 1 and the beginning of Phase 2, α_3 the ending of Phase 2 and the beginning of Phase 3, α_4 the ending of Phase 3 and the beginning of Phase 4 and α_5 the ending of Phase 5 where $\alpha_1 > \alpha_2 > \alpha_3 > \alpha_4 > \alpha_5$. Then inferences need to be made about $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$.

5. Each sherd has been assigned to one and only one Phase.

6. Within each Phase it is assumed that the ceramic production is constant and that the sherd associated with a radiocarbon date is equally likely to come from any year within that Phase. If the year of production of a sherd in the jth Phase is denoted by θ then this assumption results in a density of θ as follows

$$p(\theta|\alpha_j, \alpha_{j+1}) = \begin{cases} (\alpha_j - \alpha_{j+1})^{-1} & (\alpha_j > \theta \geq \alpha_{j+1}) \\ 0 & \text{(otherwise).} \end{cases}$$

Given these assumptions it is possible to write down the likelihood function, denoted by $l(x; \alpha)$, relating all the sixty-five radiocarbon results to the unknown parameters $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ and α_5, the beginning and ending dates of the Phases.

3 The prior information

To apply the Bayesian methodology to obtain the posterior density of $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$, we need to specify our prior belief about these parameters. In the light of no information about these values except that they are ordered we take

$$p(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = \begin{cases} 1 & (\alpha_1 > \alpha_2 > \alpha_3 > \alpha_4 > \alpha_5) \\ 0 & \text{(otherwise).} \end{cases}$$

4 The statistical inference procedures

The joint posterior density is given by

$$p(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5|x) \propto l(x; \theta)p(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$$

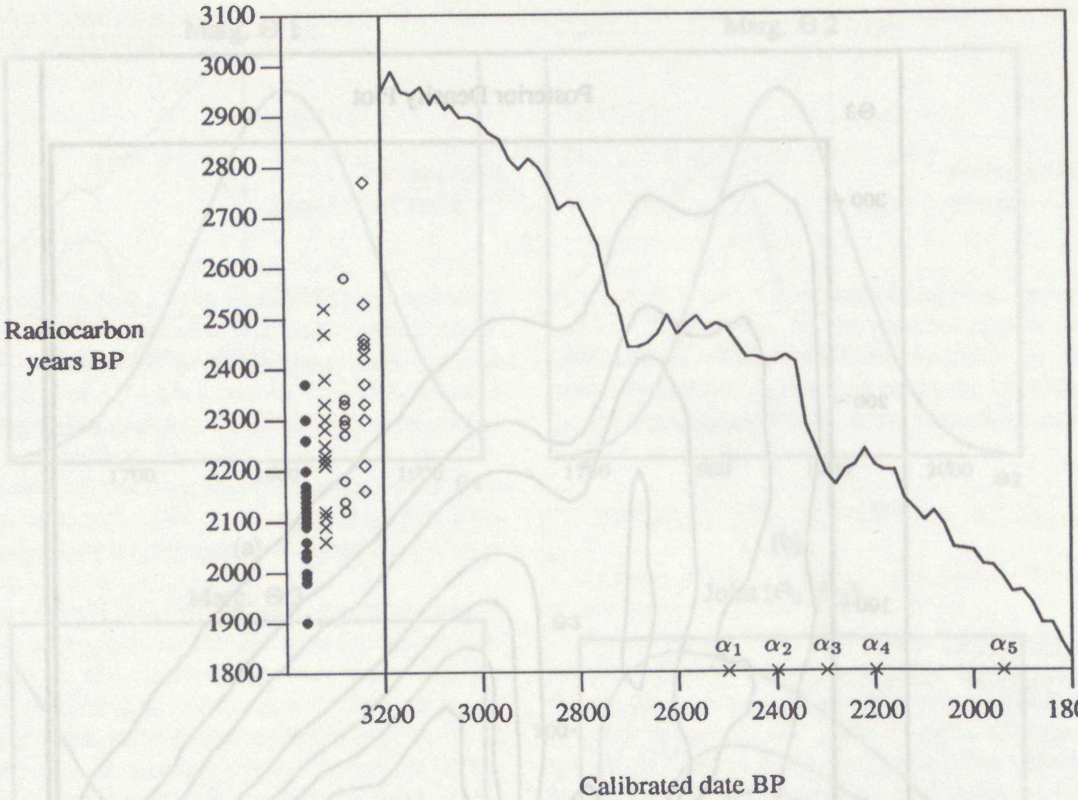

Figure 16.5: The high precision calibration curve relevant to the Danebury example. The α_i are the beginning and ending dates of the four phases on the calendar scale. The radiocarbon results for each phase are shown on the vertical axis.

This can be thought of as a surface in five dimensions, a surface that may be rather strange with various modes and ridges. To interpret it various one and two dimensional plots are extracted. This is where sophisticated software for numerical integration and contouring is required and Naylor and Smith (1988) carry this out using the BAYES4 program.

5 The interpretation of the results

The plots of the marginal densities of the α_i given by Naylor and Smith show that the posterior densities of these parameters are skew and so that reporting results in the statistically conventional way by a mean and standard deviation will be inappropriate. Furthermore plots of the joint posterior densities of the pairs α_i and α_j show that the α_i are correlated with each other. This is not unexpected as say moving the end of Phase 1 (α_2) a little earlier will have a direct influence on both α_1 and α_3, and to a lesser extent α_4 and α_5.

Despite the difficulties in interpreting such plots, which will be obvious even from the simple examples given earlier, the model nevertheless provides answers to certain specific questions which are clearly of interest to the archaeologist. Thus, although the model was set up in the first place to estimate the beginnings and endings of phases in calendar years, it also allows inferences about individual sherds to be made, namely:

- the calendar date for an archaeological context given an associated sherd from one of the phases;
- the phase of a sherd, given the calendar year;
- the phase of a sherd, given the associated radiocarbon result.

Such matters are impossible to assess from 'probability' conversions of the individual dates or indeed from any attempt to describe the data for each phase in terms of 'blanket' summary statistics.

16.6 Discussion and conclusions

There are obviously many criticisms that can and have been levelled at any attempt to model a complex situation. There are aspects in any model that both archaeologists and radiocarbon laboratories will raise objections to although we must emphasise that many of these can be overcome by some collaboration between workers in the three disciplines involved. This is clearly acknowledged by Naylor and Smith in their paper. For instance in the above study the high precision curve was not used as it was not available at that time. Another practical point is that the standard deviations of the radiocarbon dates were used as reported by the laboratory despite possible under-estimation. However in a re-analysis account of both these points has been made (Naylor 1990).

Some archaeologists will be unhappy at some of the modelling suppositions which have been put forward. For example in the case of Danebury, why are the phases non-overlapping? Why should the sherds be uniformly distributed over a Phase? Can a sherd be assigned to more than one Phase? The model can be readily adapted to encompass these and other complexities but the overall Bayesian methodology will remain the same. Of course any changes to the model would have to capture the archaeologist's perceived idea of the underlying processes generating the data. Furthermore he or she must be prepared to defend these ideas.

While there is obviously considerable scope for discussion about the adequacy of any particular model and the data to estimate its parameters, the important point is that a methodology exists which is capable of being adapted to the range of situations which archaeologists may encounter. Even if the quality and quantity of the data which is presently available may not seem to warrant the complex mathematics involved, all those concerned with radiocarbon dating must surely wish to improve matters, as can be seen in the recent intercomparison and quality control studies. To abandon the attempt to improve data treatment would be irresponsible. It should also be recognised that inferences *are* made from radiocarbon dates, so that any attempt to make inferences more objective and explicit must be to the good. Finally, although space has precluded a discussion of this topic, model building of the type described has the advantage that sensitivity analyses can be performed. That is, the effect of changes in either the data or the model can be investigated to give an idea of the range of feasible inferences which is consistent with the data.

The practical implementation of the ideas put forward here is in some ways more problematical than the development of the theory. For single radiocarbon results about which there is no specific knowledge, the CALIB program produces probabilities that are in line with the simple 'vague' prior model outlined earlier. For more complex situations involving specific prior knowledge and/or more than one date, one requires the numerical integration techniques embodied in a package like BAYES4. Alternatively the development of novel techniques based upon the Gibbs Sampler (Geman & Geman 1984) may prove easier to implement. Even if programs were more readily available, expertise will be required at least in the short term in setting up the model and interpreting the results. Nevertheless, there is no reason in principle why radiocarbon laboratories with access to modern computing facilities should not acquire this expertise once the basic statistical framework has been established.

Acknowledgements

We are grateful to Sheridan Bowman and Janet Ambers for their encouragement and for comments on a draft of this paper. We thank Caitlin Buck, who is supported by the UK Science and Engineering Research Council's Complex Stochastic Systems Initiative, for patiently producing the graphs.

Bibliography

AITCHISON, T. C., M. LEESE, D. J. MICHCZYNSKA, W. G. MOOK, R. J. OTLET, B. S. OTTAWAY, M. F. PAZDUR, J. VAN DER PLICHT, P. J. REIMER, S. W. ROBINSON, E. M. SCOTT, M. STUIVER, & B. WENINGER 1989. "A comparison of methods used for the calibration of radiocarbon dates", *Radiocarbon*, 31 (3): 846–864. 1989.

AITCHISON, T. C., B. S. OTTAWAY, & E. M. SCOTT 1987. "Statistical treatment of groups of related radiocarbon dates", *in 2nd Symp. Archaeology and 14C*. P.A.C.T., Groningen.

BOWMAN, S. 1990. *Radiocarbon Dating*. British Museum Publications, London.

CUNLIFFE, B. 1984. *Danebury: An Iron-Age Hill Fort in Hampshire*. Council for British Archaeology, Research Report 2, London.

GEMAN, S. & D. GEMAN 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6 (6): 721–741.

HELSKOG, K. & T. SCHWEDER 1989. "Estimating the number of contemporaneous houses from 14C dates", *Antiquity*, 63: 166–172.

LEESE, M. N. 1988. "Methods for finding calendar date bands from multiple-valued radiocarbon calibration curves", *in* Ruggles, C. L. N. & Rahtz, S. P. Q., (eds.), *Computer and Quantitative Methods in Archaeology 1987*, International Series 393, pp. 147–151. British Archaeological Reports, Oxford.

NAYLOR, J. C. 1990. "Personal communication".

NAYLOR, J. C. & J. E. H. SHAW 1985. *Bayes Four User Guide*. Statistics Group, Department of Mathematics, University of Nottingham, Nottingham, U.K.

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

ORTON, C. 1984. "A statistical technique for integrating C-14 dates with other forms of dating evidence", *in Computer Applications in Archaeology 1983*, pp. 115–124, Bradford. University of Bradford.

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

ROBINSON, S. W. 1986. "A computational procedure for the utilization of high-precision radiocarbon curves", Technical report, Open-File Report, United States Department of the Interior Geological Survey.

STUIVER, M. & G. W. PEARSON 1986. "High-precision calibration of the radiocarbon time scale, AD 1950–500 BC", *Radiocarbon*, 28 (2B): 805–838.

STUIVER, M. & P. REIMER 1986. "A computer program for radiocarbon age calibration", *Radiocarbon*, 28 (2B): 1022–1030.

STUIVER, M. & P. REIMER 1989. "Users' Guide to the programs CALIB and DISPLAY rev. 2.1", Technical report, Quaternary Isotope Lab, University of Washington.

VINCENT, C. H. 1988. "Treatment of discrepancies in radiocarbon dating", *Radiocarbon*, 30 (2): 157–160.

WARD, G. K. & S. R. WILSON 1978. "Procedures for comparing and combining radiocarbon age determinations: a critique", *Archaeometry*, 20 (1): 19–31.