

Bayesian interpretation of radiocarbon results

by

José Andrés Christen

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy, May, 1994.

To *all* my family

"árbol mental, frutos sabor a tiempo"

Paz (1958, p. 65).

Table of Contents

| | |
|---|-----------|
| Abstract | viii |
| Acknowledgments | ix |
| Chapter 1: Introduction | 1 |
| 1.1 History of radiocarbon dating | 1 |
| 1.1.1 Radiocarbon dating | 4 |
| 1.1.2 The need to calibrate | 5 |
| 1.2 Basic model and calibration | 9 |
| 1.3 The importance of dating | 15 |
| 1.3.1 Dating techniques | 16 |
| 1.3.2 The rôle of radiocarbon dating in archaeology | 18 |
| 1.4 Interpreting radiocarbon determinations | 20 |
| 1.5 Plan of thesis | 22 |
| Chapter 2: Bibliographical review | 24 |
| 2.1 Introduction | 24 |
| 2.2 Reliability of radiocarbon dating | 25 |
| 2.2.1 Introduction | 25 |
| 2.2.2 Overview of some interlaboratory studies | 25 |
| 2.2.3 Discussion | 29 |
| 2.3 Calibration | 30 |
| 2.3.1 Introduction | 30 |
| 2.3.2 The radiocarbon calibration data | 32 |
| 2.3.3 Early approaches for calibration | 34 |
| 2.3.4 Two major contributors | 36 |
| 2.3.5 Recent developments | 40 |

| | |
|---|------------|
| 2.3.6 Discussion | 44 |
| 2.4 Interpretation of radiocarbon determinations | 45 |
| 2.4.1 Introduction | 45 |
| 2.4.2 Early work | 45 |
| 2.4.3 Ward and Wilson (1978) | 49 |
| 2.4.4 Other major contributors | 56 |
| 2.4.5 Archaeological wiggle matching | 60 |
| 2.4.6 'El Castillo' determinations | 64 |
| 2.4.7 Bayesian approaches | 67 |
| 2.4.8 Discussion | 75 |
| 2.5 Discussion | 77 |
| Chapter 3: Calibration | 79 |
| 3.1 Modelling the calibration curve | 80 |
| 3.1.1 The process $r(\theta)$ | 81 |
| 3.1.2 Estimation of s | 87 |
| 3.2 Calibration | 93 |
| 3.3 Examples | 94 |
| 3.4 Comparison with other estimates for the variance in $r(\theta)$ | 98 |
| Chapter 4: Statistical framework | 102 |
| 4.1 Introduction | 102 |
| 4.2 Dating in archaeology | 103 |
| 4.3 The statistical framework | 106 |
| 4.3.1 Modelling approach | 106 |
| 4.3.2 Interpretation of probability within the dating problem | 110 |
| 4.3.3 Consistency | 112 |
| 4.3.4 Posterior distributions and MCMC methods | 113 |
| 4.3.5 An example | 119 |
| 4.4 Replications | 124 |

| | |
|---|------------|
| 4.4.1 Example: The Shroud of Turin | 129 |
| 4.4.2 Comment | 131 |
| Chapter 5: Examples | 134 |
| 5.1 Introduction | 134 |
| 5.2 The Chancay culture: a summarisation problem | 135 |
| 5.3 Floating chronologies | 139 |
| 5.3.1 A tree-ring chronology | 143 |
| 5.3.2 A floating chronology | 147 |
| 5.4 Contemporaneous houses | 153 |
| 5.5 Peat cores | 157 |
| 5.5.1 An example | 159 |
| 5.6 Working with phases | 166 |
| 5.6.1 A general setting | 167 |
| 5.6.2 Form of the full conditionals | 171 |
| 5.6.3 An example | 173 |
| 5.6.4 Sampling techniques | 176 |
| Chapter 6: Robustness | 180 |
| 6.1 Introduction | 180 |
| 6.1.1 Outliers in radiocarbon dating | 181 |
| 6.1.2 The summarisation problem: robustness to outliers | 183 |
| 6.1.3 Previous approaches | 186 |
| 6.2 Outliers in statistics | 192 |
| 6.2.1 What can we do? | 196 |
| 6.3 Our approach | 201 |
| 6.3.1 Shift outliers | 201 |
| 6.3.2 The extended framework | 203 |
| 6.4 Examples | 206 |
| 6.4.1 Replications | 206 |

| | |
|---|------------|
| 6.4.2 The summarisation problem: detecting outliers | 212 |
| 6.5 Remarks | 214 |
| Chapter 7: The Galgenberg site: Robust estimates | 216 |
| 7.1 Introduction | 216 |
| 7.2 The mathematical modelling of a new dating problem | 217 |
| 7.3 The Galgenberg site | 220 |
| 7.3.1 Previous approaches | 223 |
| 7.4 The dating problem | 225 |
| 7.4.1 Grouping the determinations | 227 |
| 7.4.2 The relative chronology | 228 |
| 7.4.3 Mathematical modelling | 231 |
| 7.5 Outlier identification | 235 |
| 7.6 Robust estimates | 241 |
| 7.7 Conclusions | 246 |
| Chapter 8: Discussion | 249 |
| References | 254 |

Abstract

Over the last thirty years radiocarbon dating has been widely used in archaeology and related fields to address a wide-range of chronological questions. Because of some inherent stochastic factors of a complex nature, radiocarbon dating presents a rich source of challenging statistical problems. The chronological questions posed commonly involve the interpretation of groups of radiocarbon determinations and often substantial amounts of *a priori* information are available. The statistical techniques used up to very recently could only deal with the analysis of one determination at a time, and no prior information could be included in the analysis. However, over the last few years some problems have been successfully tackled using the Bayesian paradigm. In this thesis we expand that work and develop a general statistical framework for the Bayesian interpretation of radiocarbon determinations.

Firstly we consider the problem of radiocarbon calibration and develop a novel approach. Secondly we develop a statistical framework which permits the inclusion of prior archaeological knowledge and illustrate its use with a wide-range of examples. We discuss various generic problems some of which are, replications, summarisation, floating chronologies and archaeological phase structures. The techniques used to obtain the posterior distributions of interest are numerical and, in most of the cases, we have used Markov chain Monte Carlo (MCMC) methods. We also discuss the sampling routines needed for the implementation of the MCMC methods used in our examples. Thirdly we address the very important problem of outliers in radiocarbon dating and develop an original methodology for the identification of outliers in sets of radiocarbon determinations. We show how our framework can be extended to permit the identification of outliers. Finally we apply this extended framework to the analysis of a substantial archaeological dating problem.

Acknowledgments

I am very grateful indeed to my supervisor Dr Cliff Litton for his patience, support, encouragement and enlightening comments throughout the development of my work. Also to Caitlin Buck for her invaluable comments and corrections to improve both the contents and presentation of my work. For both of them I cannot find the correct words to say *gracias*. I must thank as well the archaeologists James Kenworthy who helped me understand the archaeological characteristics of the example developed in Chapter 7.

I am very grateful as well to all of you who helped me improve the English and style in several parts of the thesis: Damian Clancy, Jon Foster, Terry Watts, Marc Kennedy and Dr Bob Laxton. Also I most thank the Nottingham University statistics research group for sharing their rich intellectual environment with me. This research was supported by the Dirección General de Asuntos del Personal Académico, Universidad Nacional Autónoma de México.

Words will hide as always although this time they seem to vanish; please forgive me if I omit a name that should have been mentioned or for not saying all that I should say. To all of you, so far and so near, and specially to Juan Carlos, Jack, Michelle, Mary, Mery, tía Chata, Mum and Dad, thank you for your letters and love. And to all of you, my dearest friends, Caitlin, Mark, JC, Rom, Bibi, Pilar, Mónica, Stefano, Tomás, Adrian and Sofia, thank you loads for being there.

Chapter 1

Introduction

1.1 History of radiocarbon dating

"Nuclear physical data indicate that cosmic-ray neutrons produce ^{14}C and ^3He from atmospheric nitrogen, the radiocarbon being the principal product. The purpose of this letter is (...) to suggest that radiocarbon might be found in living matter (...)"

Hidden among hundreds of volumes in the store of the University of Nottingham Science Library, one can find this extract in a two page letter sent by W. Libby in 1946 to the journal *Physics Review* (Libby 1946). It contains a theoretical analysis that predicted the content of radiocarbon in all living matter. Libby conjectured that cosmic rays, when entering the earth's atmosphere, split nitrogen into Helium-3 and Carbon-14, ^{14}C (radiocarbon). The ^{14}C is then incorporated into the biosphere in the form of carbon dioxide, and absorbed first by plants for photosynthesis and consequently by all living matter in the food chain. He postulated that by this process, all living matter would eventually contain radiocarbon.

Carbon has three naturally occurring isotopes, ^{12}C , ^{13}C and ^{14}C . Of these, only ^{14}C is radioactive. This means that ^{14}C is an unstable atom and eventually will decay (by emission of an electron) to Nitrogen-14. This decay process follows the 'law of radioactive decay'

$$M = M_0 \exp\left(-\frac{\ln 2}{T_{1/2}} t\right) \quad (1.1)$$

where M is the $^{14}\text{C}/^{12}\text{C}$ ratio remaining after a time t , having started with a $^{14}\text{C}/^{12}\text{C}$ ratio of M_0 , and $T_{1/2}$ is called the 'half-life'. The 'mean life' is given by

$\tau = \frac{T_{1/2}}{\ln 2}$. For ^{14}C , it is currently estimated that $T_{1/2} = 5730$ years but, for the purpose of radiocarbon dating Libby's value of $T_{1/2} = 5568$ years is used, with the corresponding mean life of $\tau = 8033$.

If we suppose that ^{14}C is created by cosmic rays at a constant rate and since it is constantly decaying, as seen above, we can expect the ^{14}C in the biosphere to remain at an equilibrium level. In the 1946 letter to *Physics Review*, Libby obtained an estimate of such an equilibrium level based on some theoretical considerations. Libby proceeded to validate his predictions using experimental observations, and one year later he and his colleagues published a paper in the journal *Science* explaining that (Anderson *et al.* 1947),

"It has recently been suggested [Libby (1946)] (...) that (...) all carbon in living matter (...) should be radioactive to the extent of 10 disintegrations/minute/gram. In view of the 5000-year half-life of radiocarbon (...) it was further expected that it should be absent from such geologically "old" carbon sources as petroleum, coal, or limestone. These predictions were investigated by examining the radiocarbon activity of two series of isotopically enriched samples of methane. The first series was derived from petroleum methane (referred as petromethane) and the other from the Patapsco Sewage Plant of the city of Baltimore (referred as biomethane). Measurements on the enriched biomethane samples established the activity of "living" carbon to be 10.5 disintegrations/minute/gram, in good agreement with the predicted value. On the other hand, enrichment of petromethane by a factor of 25 failed to show activity beyond the limits of experimental error, in line with the theory that cosmic rays produce our activity."

Thus the basis for radiocarbon dating had been established. When living matter dies, it is removed from the biosphere and consequently its ^{14}C starts decaying without being replaced by more ^{14}C atoms from the environment. The $^{14}\text{C}/^{12}\text{C}$ ratio in an organism at death (M_0 in Equation 1.1) should be approximately the same as the equilibrium $^{14}\text{C}/^{12}\text{C}$ ratio in the atmosphere. By measuring the $^{14}\text{C}/^{12}\text{C}$ ratio in an 'old' object (M in Equation 1.1) containing organic matter, one can use the 'law of radioactive decay' to estimate its age, t , which is given by

$$t = -\tau \ln(M/M_0).$$

(In practice the $^{13}\text{C}/^{12}\text{C}$ ratio present in a sample has to be accounted for when calculating the $^{14}\text{C}/^{12}\text{C}$ ratio, see Gillespie 1986, chapter 4, for details.)

The above ideas were first proposed by Libby who also proceeded to demonstrate that the distribution of ^{14}C could be considered to be uniform around the world (although very recently some doubts about this global uniformity have emerged, see Pearson and Stuiver 1993). He continued the research into radiocarbon dating and in 1949 he and his colleagues published the first paper containing dates produced by this method (Libby *et al.* 1949),

"Having established the world-wide uniformity of the radiocarbon assay at the present time, it seems logical assumption that this would have been true in ancient times. Assuming this (...) one can calculate the specific [^{14}C] activity to be expected at any given time interval elapsed since the removal of any carbonaceous material from equilibrium with the life cycle. For living materials this probably coincides with the time of death (...)"

The first two items to be dated were wood samples from two Egyptian tombs of 'well established age'. The estimated ages resulted in good agreement with the expected ages, confirming the possibility of being able to date organic matter using radiocarbon. A purely theoretical consideration of cosmic radiation had evolved into a practical means for age determination.

Since 1949 radiocarbon dating has been used intensively, principally within archaeological research (other applications of radiocarbon dating can be found in geology, hydrology, earth and environmental sciences among other disciplines). The technique has been refined and has evolved into a highly sophisticated process involving both chemical and physical procedures. In the next Section we briefly review the techniques and the processes involved in radiocarbon dating. However, it is not our intention to explain the processes in great detail. A more

detailed exposition can be found in Bowman (1990) or Gillespie (1986).

1.1.1 Radiocarbon dating

As seen above, the basic principle of radiocarbon dating is to measure the proportion of ^{14}C in the object to be dated and compare it with the 'equilibrium level' in the atmosphere in order to obtain an approximation to its age, using the law of radioactive decay.

The objects to be dated need to contain organic matter. The most commonly dated objects or materials are wood, charcoal, seeds, bone, sediments and peat. Because the half-life of radiocarbon is around 5000 years the effective range for radiocarbon dating is from 200 to 25,000 years, making it especially useful for archaeology. The process of radiocarbon dating basically consists of measuring the ratio $^{14}\text{C}/^{12}\text{C}$ in the organic matter contained in the object to be dated. Letting this ratio be M , we have that the age for the object is

$$t = -\tau \ln(M/M_0), \quad (1.2)$$

where M_0 is the equilibrium level in the atmosphere and τ ($= 8033$) is the 'mean life' of ^{14}C as defined above. In radiocarbon dating M_0 is a constant known as the 'modern standard' (see Gillespie 1986, p. 21).

There are two ways of measuring the $^{14}\text{C}/^{12}\text{C}$ ratio, the indirect method and the direct method, the latter being the most common one. The indirect method ('conventional radiocarbon dating') consists of measuring the electrons emitted when the ^{14}C atoms decay. The measurements typically take between 10 and 72 hours, and from that the $^{14}\text{C}/^{12}\text{C}$ ratio is measured up to some precision (the 'counting error'). The direct method use the technique of 'accelerator mass spectrometry' ('AMS radiocarbon dating'), which directly measures the amount of

^{14}C atoms in the sample, up to some precision (the 'counting error'). This is a new technique which has only been generally available since 1986. More details about both conventional and AMS radiocarbon dating can be found in Aitken (1990, chapter 4).

For both conventional and AMS radiocarbon dating, the carbon contained in the objects to be dated (charcoal, wood, seeds, bone, etc.) needs to be isolated. This involves physical and chemical processes that may increase the imprecisions in the measuring of the $^{14}\text{C}/^{12}\text{C}$ ratio. Combining these imprecisions with the counting error, the radiocarbon dating laboratories calculate an error, or standard deviation, for the estimation of the $^{14}\text{C}/^{12}\text{C}$ ratio from which they calculate the object's age with an error or standard deviation (see Bowman 1990, chapter 3, for a more detailed discussion).

Thus, the final result of a radiocarbon dating process is a radiocarbon determination that consists of an estimated 'radiocarbon year' y and an error, or standard deviation σ , reflecting the uncertainty in the process. In this thesis we will write radiocarbon determinations in the form $y \pm \sigma$, for example 1400 ± 70 , and this notation should cause no confusion. The meaning of a 'radiocarbon year' will be explained in the next Section.

1.1.2 The need to calibrate

Normally it is not possible to obtain the precise date for objects found in archaeological or geological excavations and, for this reason, it is difficult to assess the accuracy of the radiocarbon dating technique. Nevertheless, by counting the rings in trees and measuring their widths (a technique known as 'dendrochronology') one can obtain samples of wood of known age. By radiocarbon dating such samples it is possible to evaluate the accuracy of radiocarbon

determinations. In fact, by dating wood samples of known age, differences between the radiocarbon and the true ages soon became apparent. Thus it was clear that there was a problem with the technique, the problem being related to the assumption that the equilibrium $^{14}\text{C}/^{12}\text{C}$ ratio in the atmosphere had remained constant through time. Subsequent research has shown that cosmic radiation varies through time and is affected by factors like sunspots, variations in geomagnetism and the like (a review on the subject can be found in Damon *et al.* 1989). Consequently the atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio varies through time.

Since the initial $^{14}\text{C}/^{12}\text{C}$ ratio in an organism (that is the $^{14}\text{C}/^{12}\text{C}$ ratio in an organism at death) is considered to be approximately the same as the atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio, M_0 in Equation 1.2 will not be constant as it is dependent on the atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio existing when the sample died. Indeed, we do not know the corresponding M_0 for all samples and thus an approximate value is used, the 'modern standard', to calculate the 'radiocarbon age' of the samples (t in Equation 1.2). This radiocarbon age will then differ from the true (calendar) age of the samples depending on how much their initial $^{14}\text{C}/^{12}\text{C}$ ratio differed from the 'modern standard'. This is the reason why radiocarbon determinations are measured in 'radiocarbon years' since they generally differ from calendar years. The process of transforming from radiocarbon years to calendar years is a problem of statistical calibration and is known as 'radiocarbon calibration'.

Over the last 20 years the international radiocarbon community has invested much effort in dating wood samples of known age so as to relate radiocarbon years to calendar years. However, until 1986 there were only low precision results and much discussion took place concerning the general applicability of the data and other technical details (see Section 2.3 for a bibliographical review of the subject). Most of the difficulties have now been overcome and definite results have been obtained by Pearson and Stuiver (1986), Stuiver and Pearson (1986) and

Pearson *et. al.* (1986). The results of these three papers form the basis of what the radiocarbon community calls ‘the internationally agreed high-precision calibration curve’. This consists of approximately 460 dendrochronologically determined calendar years t_k ’s spaced approximately every 20 years from 7200 BC to 1950 AD, each one associated with a set of replicated radiocarbon determinations that give the average result $x_k \pm \sigma_k$. We call $[t_k, (x_k \pm \sigma_k)]$ a ‘calibration observation’ and the whole of the vector $z = ([t_1, (x_1 \pm \sigma_1)], \dots, [t_n, (x_n \pm \sigma_n)])$ the ‘high-precision calibration data’. Typically the σ_k ’s have values in the range of 10 to 20 years. These values are small in comparison to the standard deviation of a typical radiocarbon determination, which is in the range of 40 to 80 years. This is the reason why z is called high-precision calibration data.

The t_k ’s above and, unless clearly stated, all dates (and variables representing dates) are measured in ‘years before present’, which are the number of years before 1950 AD. ‘years BP’ (Before Present) is then used when referring to calendar years, and ‘years bp’ when referring to radiocarbon years (this is a standard notation in radiocarbon and archaeology). When we say 5,500 calendar years BP (or radiocarbon years bp) one can think of it as $-5,500$ years with year 0 BP (bp) equal to 1950 AD ie. 3,551 BC since year ‘0’ does not exist. Therefore the t_k ’s above are measured in years BP, since they represent calendar years, and the x_k ’s in years bp, since they represent radiocarbon years. In all of the plots we present, the ‘x’ axis corresponds to the calendar years but inverting its direction. As a result of this, older (ie. large) years BP are to the left of the plot and younger (ie. small) years BP are to the right (see Figures 1.1, 1.2, 1.3 or 1.4). Radiocarbon years are plotted in the ‘y’ axis (with its usual direction). This layout has been adopted by radiocarbon laboratories and archaeologists for some years and we will use it in this thesis. To illustrate these ideas we have plotted a portion of the high-precision calibration data in Figure 1.1.

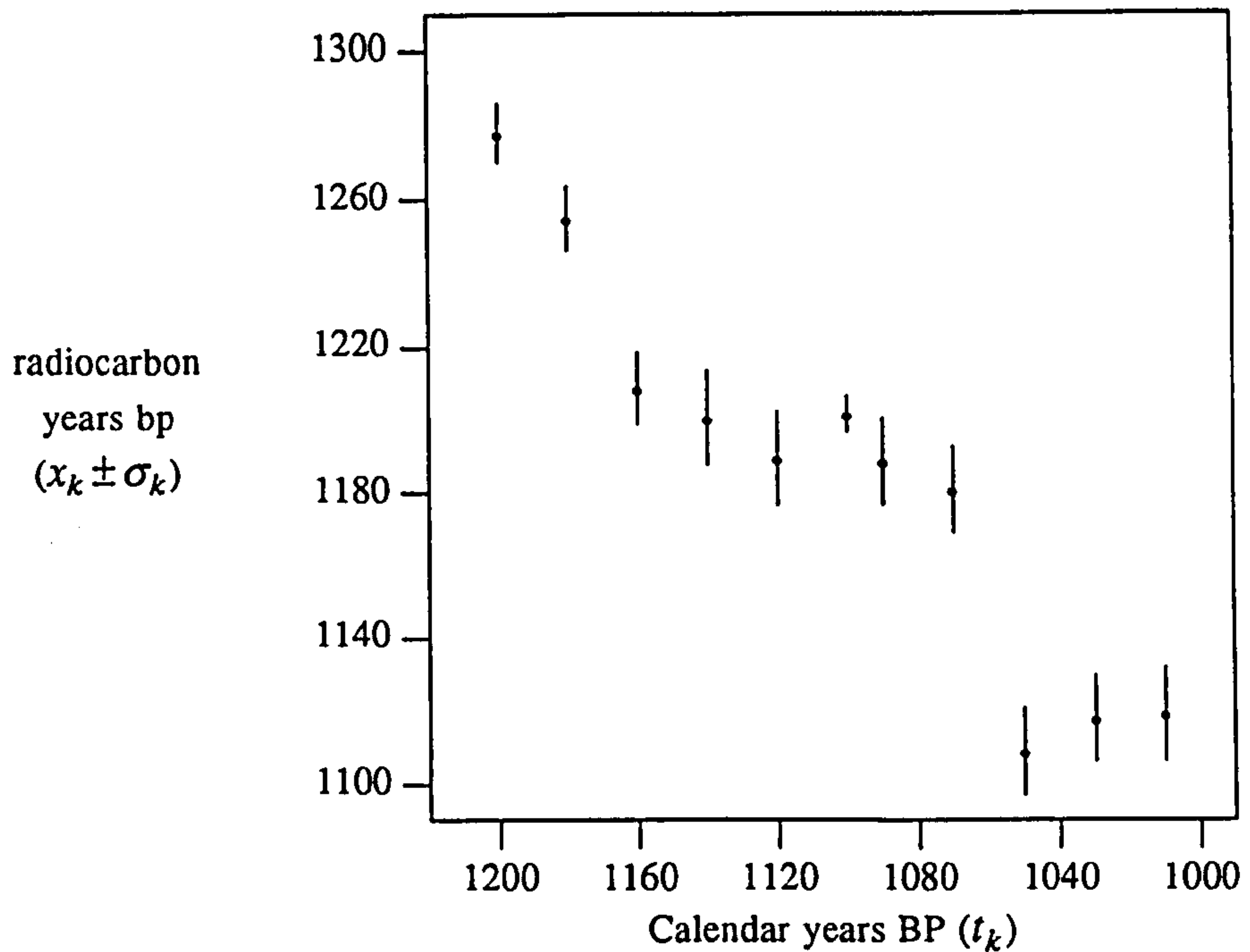


Figure 1.1

Section of the high-precision calibration data.

The dots represent the (t_k, x_k) points
and the lines are from $(t_k, x_k - \sigma_k)$ to $(t_k, x_k + \sigma_k)$.

There is still discussion about how the high-precision calibration data should be used to create a calibration curve so that for every calendar year we could obtain its corresponding radiocarbon year. One possibility is to smooth the data using the variances σ_k^2 's, but a common practise is to neglect the σ_k^2 's and use a piece-wise linear approximation, passing through the points (t_k, x_k) 's. This latter method results in a simple-to-use and, since the standard deviations of Pearson and Stuiver's data are relatively small, a reasonably accurate calibration curve. (This subject will be discussed in more detail in Section 2.3 and in Chapter 3.)

Throughout this thesis we denote the piece-wise linear calibration curve by $\mu(\cdot)$ which is defined as

$$\mu(\theta) = x_k \left(\frac{\theta - t_{k-1}}{t_k - t_{k-1}} \right) + x_{k-1} \left(\frac{t_k - \theta}{t_k - t_{k-1}} \right); \quad t_k > \theta \geq t_{k-1}, \quad k = 1, 2, \dots, n, \quad (1.3)$$

letting $t_0 = 0$ and $x_0 = 0$. A plot showing the piece-wise linear calibration curve, $\mu(\theta)$, for approximately the last 9000 years, is given in Figure 1.2. The actual calibration process will be introduced in the next Section where we study the basic statistical model for the radiocarbon determinations.

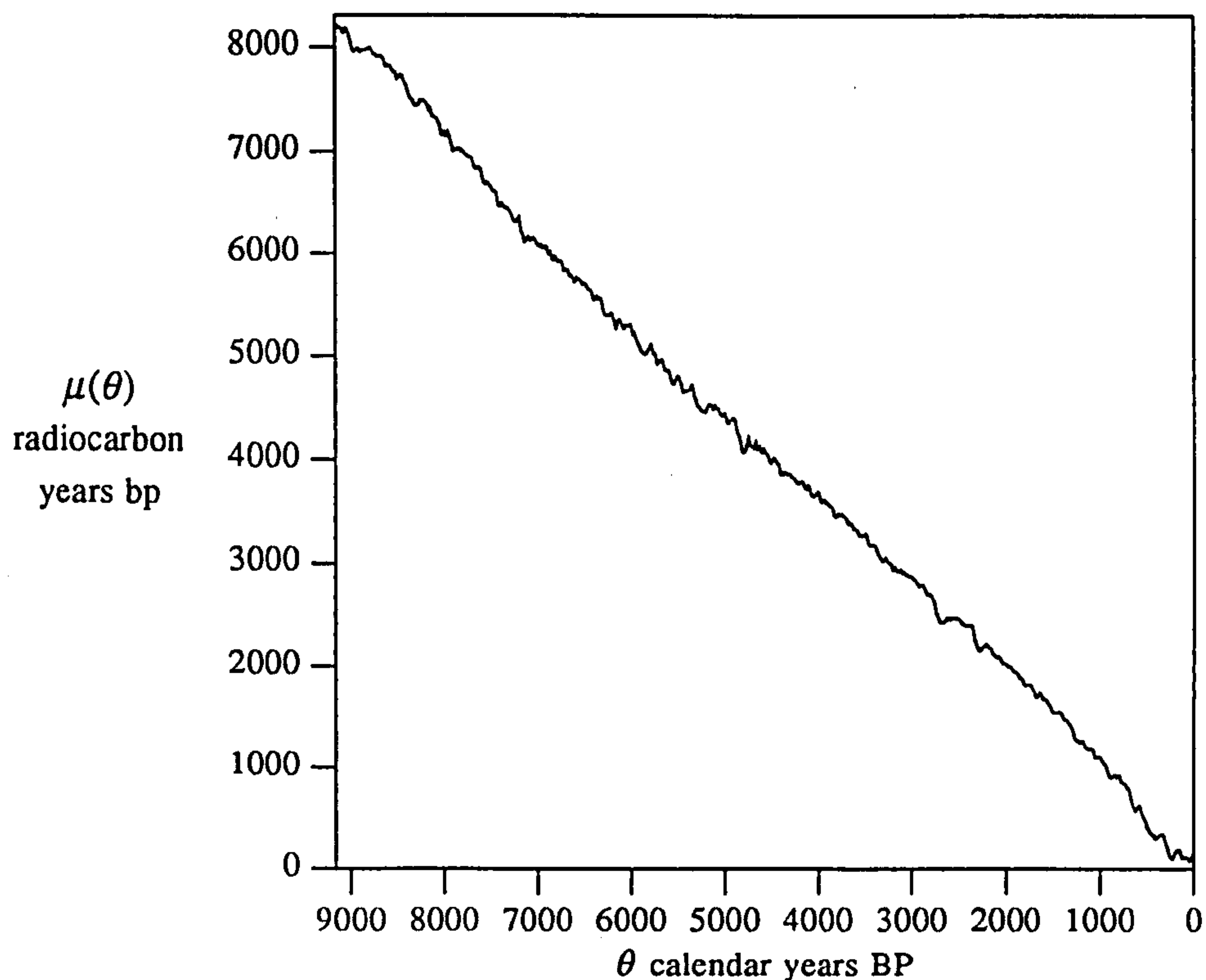


Figure 1.2
Piece-wise linear calibration curve $\mu(\theta)$
calculated using the high-precision calibration data.

1.2 Basic model and calibration

Suppose we have a radiocarbon determination $y \pm \sigma$ associated with an unknown calendar year θ . The calendar year θ corresponds to the year in which the organic material contained in the dated object died and is what we call 'the calendar year associated with the object'. As we said above, transforming from radiocarbon years to calendar years is known as radiocarbon calibration.

Therefore, within the Bayesian framework, radiocarbon calibration consists of finding $f(\theta \mid y, \sigma)$, the posterior distribution of θ given the determination $y \pm \sigma$.

The most common and widely accepted assumption in the statistical analysis of radiocarbon determinations is that they are normally distributed, having the 'true' radiocarbon age as mean and the reported laboratory error as standard deviation. This assumption is made (explicitly or implicitly) by the vast majority of the researchers in the field (see Libby 1954, Ward and Wilson 1978, Litton and Leese 1991) and we will follow this convention. In other words, suppose μ is the 'true' radiocarbon age of the dated object, then

$$y \mid \mu, \sigma \sim N(\mu, \sigma^2).$$

The above normal model has been assumed because the radiocarbon determinations arise by counting the number of ^{14}C atoms decaying in a period of time. This will have a Poisson distribution which can be approximated by a normal distribution. However, at the present time, the error in a radiocarbon determination is no longer based entirely on the 'counting error', but on other factors as well. The way this error is calculated depends somewhat on the radiocarbon laboratory, but it is usually carried out by a process of adding error factors. The usual statistical practise is to assume these additional errors are normal and so the overall error will be normal. There is no clear experimental evidence to suggest that such an assumption is not reasonable and therefore the normal errors model is generally accepted.

However, if we find a strong reason to believe that the normal model is incorrect, it could be substituted by, for example, a heavy-tailed distribution like the 't' distribution or any other distribution considered suitable. This will obviously affect any statistical technique used in the interpretation of radiocarbon determinations, including those presented in this work. Nevertheless, due to the

flexibility of the statistical tools adopted here, the techniques presented could be adapted to work with a different choice of a basic model. We do not, however, intend to explore this possibility within this thesis.

Within the Bayesian framework the basic model for radiocarbon determinations has been established following similar ideas to those presented above (see Naylor and Smith 1988, Litton and Leese 1991). If we have a radiocarbon determination $y \pm \sigma$ associated with the calendar year θ , we assume that

$$y \mid \theta, \sigma \sim N(\mu(\theta), \sigma^2),$$

where $\mu(\theta)$ is the piece-wise linear calibration curve. Therefore the model states that a radiocarbon determination is normally distributed with mean equal to $\mu(\theta)$, the ^{true} radiocarbon age corresponding to year θ , and variance σ^2 , which represents the experimental errors. Given the above model we assume σ to be known and, given a prior distribution for θ , we calculate $f(\theta \mid y, \sigma)$ using Bayes' theorem.

As mentioned above, the standard deviation, σ , reported for a radiocarbon determination $y \pm \sigma$ is evaluated by the radiocarbon laboratory theoretically as well as empirically and, strictly speaking, depends on the observed radiocarbon year y . It is not rigorously correct to assume it to be known. However, it is not at all clear what other approach could be followed. Furthermore, statistical articles on radiocarbon dating generally use this approach: see Naylor and Smith (1988), Clark (1979), Ward and Wilson (1978), Law (1975), Aitchison *et al.* (1989). Thus, at present, assuming σ to be known seems to be the only practical and viable way to proceed.

To facilitate notation we are going to avoid explicitly conditioning on σ and simply write

$$y \mid \theta \sim N(\mu(\theta), \sigma^2). \quad (1.4)$$

This convention will be followed throughout the thesis.

Now, using a vague prior distribution for θ , the resulting posterior distribution is

$$f(\theta | y) \propto \exp\left\{-\frac{(y-\mu(\theta))^2}{2\sigma^2}\right\}. \quad (1.5)$$

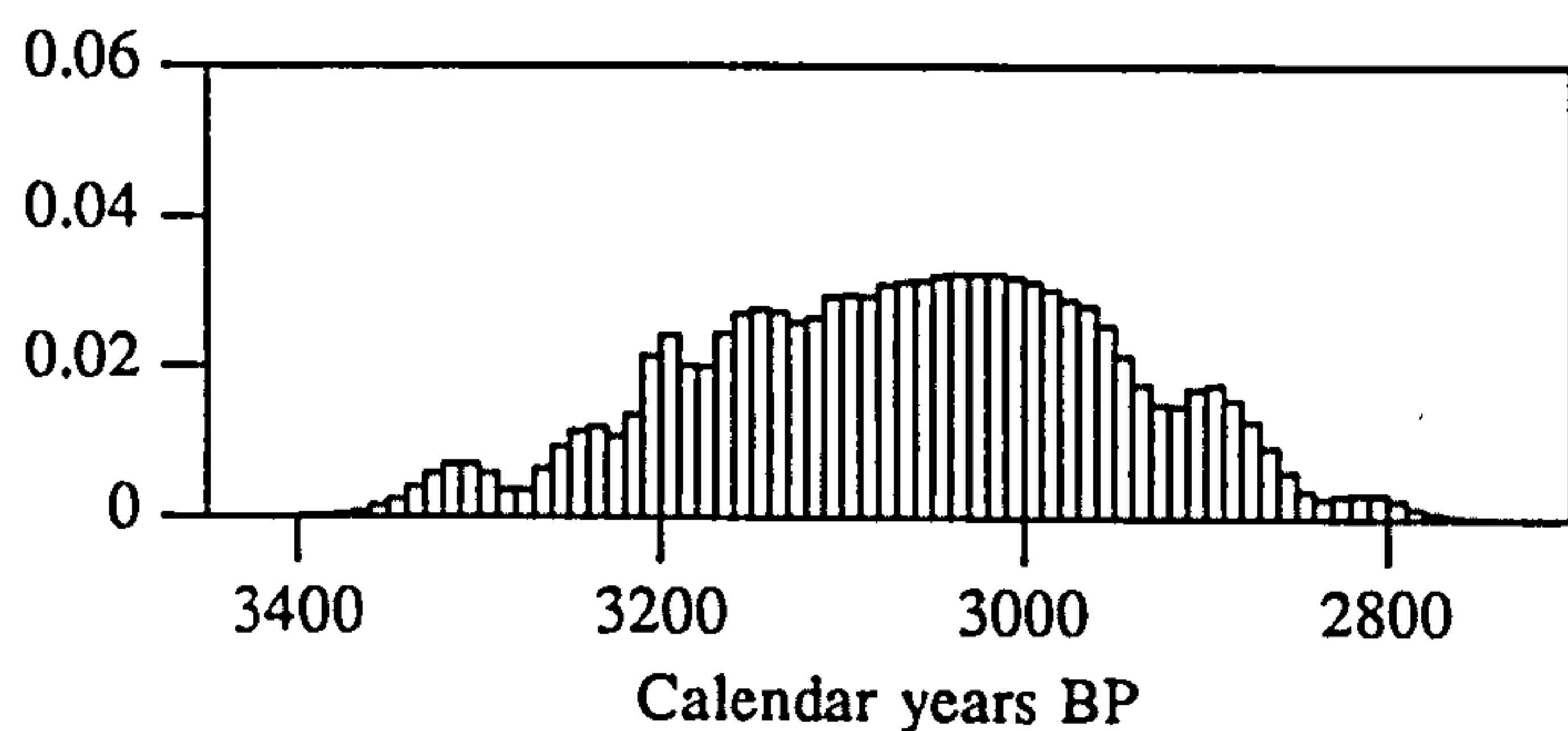


Figure 1.3
Histogram showing the posterior distribution on the calendar scale
for radiocarbon determination 2900 ± 80 .

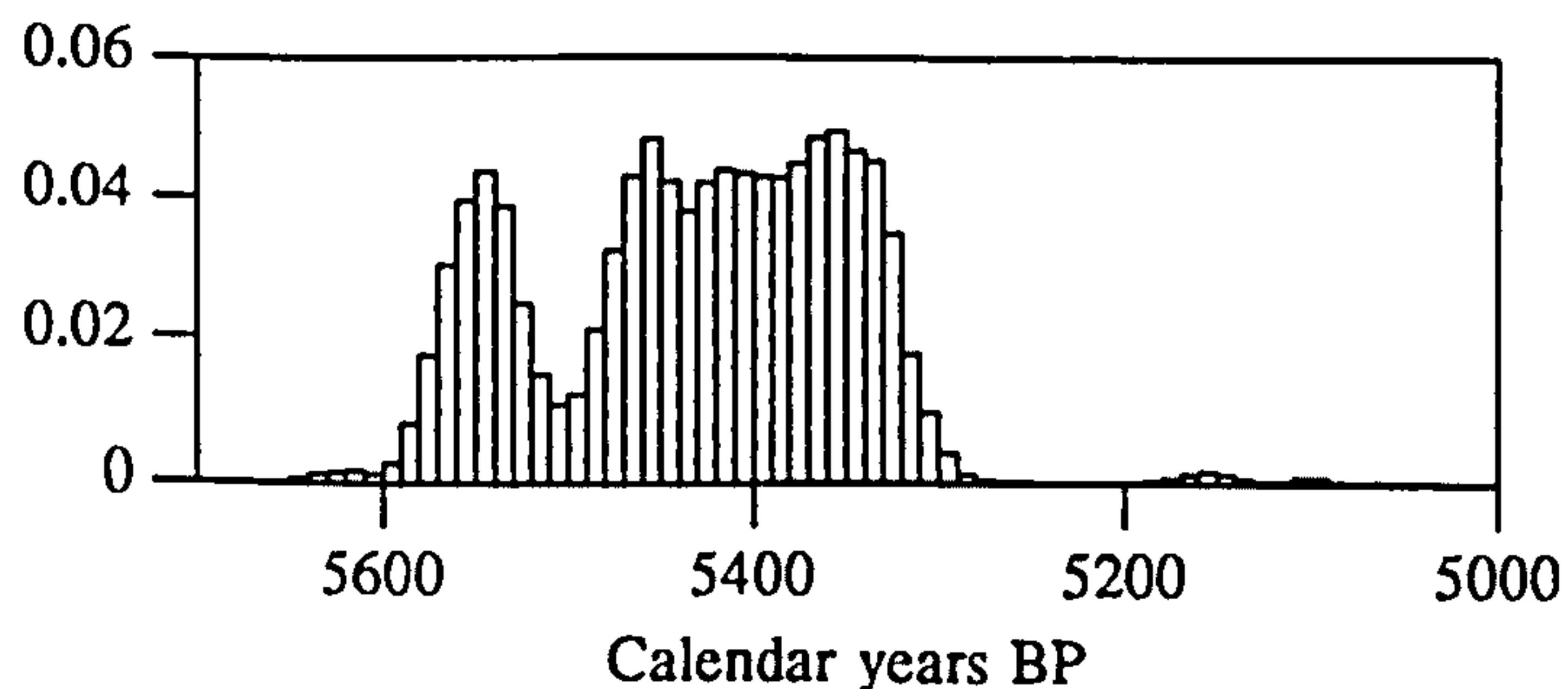


Figure 1.4
Histogram showing the posterior distribution on the calendar scale
for radiocarbon determination 4700 ± 60 .

Given the definition of the piece-wise linear calibration curve, $\mu(\theta)$, it is necessary to calculate $f(\theta | y)$ using numerical methods. Histograms showing examples of such distributions can be found in Figures 1.3 and 1.4. We have chosen two determinations from different parts of the calibration curve, 2900 ± 80

and 4700 ± 60 . The magnitude of the standard deviations are typical of the errors one would expect from a routine radiocarbon analysis.

The distributions $f(\theta | y)$ tend to have uncommon (to statisticians) shapes. Usually they have peaks or flat areas with not much regularity or symmetry. This results in multimodal distributions that differ greatly from the usual distributions used in other areas of statistics. These 'unpleasant' features of the distributions $f(\theta | y)$ are caused by the non-monotonic nature of the calibration curve $\mu(\theta)$. A powerful tool for understanding why, is to examine the derivative of $f(\theta | y)$, that is

$$f'(\theta | y) \propto \frac{(y - \mu(\theta))}{\sigma^2} \mu'(\theta) \exp \left\{ -\frac{(y - \mu(\theta))^2}{2\sigma^2} \right\}.$$

Note that the sign of this derivative is given by the term $\frac{(y - \mu(\theta))}{\sigma^2} \mu'(\theta)$. If $\mu(\theta) > y$ the sign of the derivative is the sign of $-\mu'(\theta)$ and thus $f(\theta | y)$ is increasing when $\mu(\theta)$ is decreasing and vice versa. When $\mu(\theta) < y$, $f(\theta | y)$ increases and decreases with $\mu(\theta)$. Indeed, the global maximums of $f(\theta | y)$ are at $\mu(\theta) = y$ (the 'crossings'). $f'(\theta | y)$ does not exist at the knots of the calibration data since $\mu(\theta)$ is not smooth there. This is reflected with unsmooth 'spikes' in $f(\theta | y)$.

The above analysis can be well illustrated with a plot of a particular $f(\theta | y)$ and the relevant part of $\mu(\theta)$. In Figure 1.5 we do this with the same determination (4700 ± 60) as for Figure 1.3. We have calculated the actual density $f(\theta | y)$ and not a histogram-like approximation, otherwise the fine details of $f(\theta | y)$, that we are trying to observe, would be smoothed out in the histogram's ten year bins. (Note, however, that the radiocarbon community prefers to use histograms since they are easier to understand and interpret by non-statisticians.)

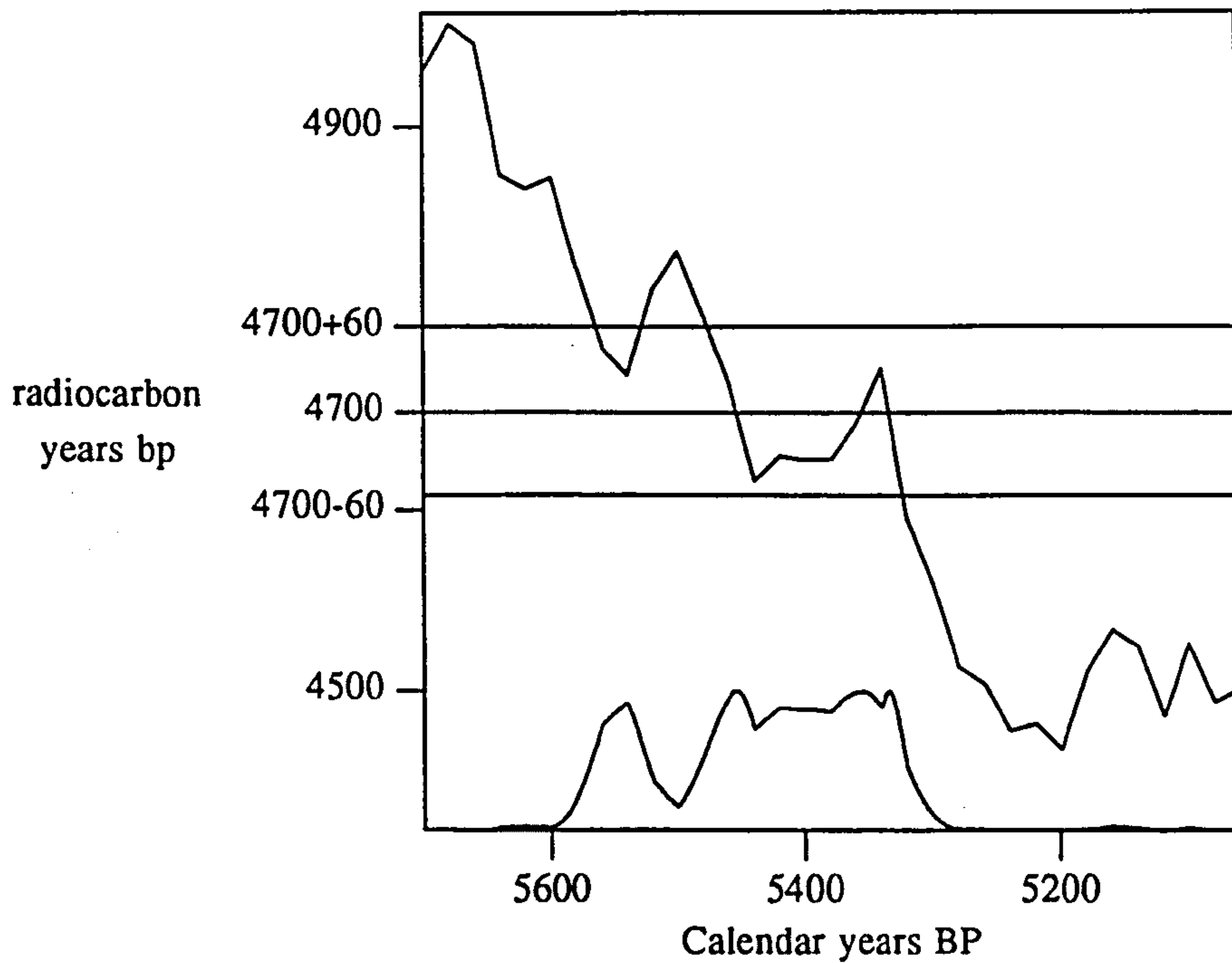


Figure 1.5
Posterior distribution in the calendar scale
for radiocarbon determination 4700 ± 60 and
the relevant part of the piece-wise linear calibration curve.

From Figure 1.5 we see that before 5450 BP the distribution $f(\theta | y)$ increases and decreases inversely with $\mu(\theta)$, giving a 'mirror' image of $\mu(\theta)$. Note that if $\mu(\theta)$ is approximately constant on a region, $(y - \mu(\theta))$ will be approximately constant as well and this will be reflected with a 'flat' region in $f(\theta | y)$, as for 5420 to 5380 BP in Figure 1.5. For the rest of the calendar scale, where $\mu(\theta) < y$, $f(\theta | y)$ simply 'mimics' the calibration curve and, of course, damps to zero as $(y - \mu(\theta))$ increases.

It is difficult to interpret distributions like the ones shown in Figures 1.3 and 1.4 since θ represents a fixed year in the past. What does a probability distribution for θ mean in such context? This problem has been one of the major sources of confusion in radiocarbon dating. The basis of our (Bayesian) interpretation is

the following. We do not believe θ to be intrinsically random in any sense, rather we take the probability distribution of θ to represent our up-to-date knowledge about the year θ . Thus it is not the year θ , but our knowledge about such a year that is uncertain. The uncertainty is then measured using the probability distribution $f(\theta | y)$.

Good ways of summarising such distributions $f(\theta | y)$ are difficult to envisage. Point estimates like the ‘maximum *a posteriori*’ estimator (MAP estimator, the maximum of $f(\theta | y)$) are obviously inappropriate due to the multimodality and asymmetry of the distributions. Highest posterior density sets (HPD sets) can be of some use but, in general, our experience tell us that the histograms themselves are the most adequate tools for interpreting the distributions $f(\theta | y)$. Indeed, a good understanding of histograms and density functions (and probability) is needed before histograms like the ones in Figures 1.3 and 1.4 can be interpreted.

1.3 The importance of dating

In this thesis we will be mostly interested in the interpretation of radiocarbon determinations in archaeology. Ultimately our aim will be to develop sound statistical techniques to date a wide variety of archaeological ‘phenomena’, using radiocarbon determinations. We may ask ourselves, however, why should we be interested in dating in the first place? In this and subsequent Sections we will address this question explaining the importance of dating to archaeology. We will then mention the different dating techniques used in archaeology and within that the important rôle of radiocarbon dating. We do this not only to motivate our work, but to give us a perspective on the relevant factors to be considered in the interpretation of radiocarbon determinations.

Measuring time is one of the most important components in the organisation of any society. Meetings, schedules, news, elections, history are almost unthinkable for us without the presence of accurate and reliable measures of time. Much the same is true in archaeology. The organisation of knowledge about events in the past is fundamentally based on the measuring of time, the time-elapsed or precise dates. This has been accurately explained by Renfrew (1973, p. 21),

"Dating is crucial to archaeology. Without a reliable chronology the past is chaotic: there is no way of relating or ordering people, events and cultures into the coherent narrative which the prehistorian seeks to construct."

Indeed, dating is one of the most crucial parts of archaeology, and of all the dating techniques available, radiocarbon dating is the most widely used. Below we briefly review the most common dating techniques and how they are used in archaeology.

1.3.1 Dating techniques

Archaeologists use a series of tools and techniques for measuring time and obtaining dates. Throughout the development of archaeology several dating techniques have been used. Before the 19th century, the understanding of the past (as with many other parts of human activity) was based on theology. A prime example is the work of Archbishop Ussher who used the genealogical records in the Bible to date the 'Creation' to October 23th 4004 BC.

With the development of rationalism, new approaches to dating appeared. These new approaches were more concerned with the logical organisation of the archaeological evidence available and less with theological dogma. Archaeologists used historical records and a progression of logical reasoning to date events in the past. By a detailed cross-matching of recorded events with calendars of

ancient civilisations, European archaeologists of the 19th and first half of the 20th centuries constructed a chronology for the past (Renfrew 1973, p. 27),

"Until the discovery of radiocarbon dating (...) there was really only one reliable way of dating events in European prehistory (...) This was by the early records of the great civilizations, which extended in some cases as far back as 3000 BC. Before that, there were no written records anywhere."

Indeed, only the tiny proportion of prehistoric events described in inscriptions or records of ancient civilisations can be accurately dated. For the rest of prehistory, dating was still a matter of broad estimates or, more likely, pure speculation.

As more systematic techniques for excavation were developed, stratigraphic relationships between contexts on archaeological sites were identified and chronological relationships began to be identified. The stratigraphic relationships on a site can provide us with very reliable chronological orderings for events and contexts in the past. Such chronologies are, however, only relative. That is, with stratigraphic relationships we may observe that 'A is before B' and 'B is before C', but very rarely can we obtain any estimates for the absolute position of A, B or C in time. Another problem of stratigraphy is that it is normally restricted to a single archaeological site and, therefore, using stratigraphy alone to build more complex chronologies is virtually impossible (see Orton 1980, p. 60 for further discussion).

Apart from stratigraphy, archaeologists can make use of other sources of information to obtain chronologies. These include the cross-matching of stylistic characteristics in pottery and the comparison of typological features in tools or weapons etc. (see Orton 1980 p. 81-88, Renfrew 1973 p. 40-52 or Fedick and Taube 1991 for examples of this). Again, such chronologies tend to be only relative and restricted to particular ancient cultures or archaeological sites.

The above dating techniques were the only ones available to archaeologists until the development of nuclear physics in the 1940's. With these advances in science came the invention of radiocarbon dating and other science-based dating techniques. Apart from radiocarbon dating, there are other dating techniques that use scientific principles to date objects. These include, potassium-argon dating, uranium-series dating, fission-track dating, thermoluminescence dating etc. (see Aitken 1990 for a review of these subjects). Also, there are other dating techniques based on the scientific study of environmental changes, these include dendrochronology, pollen records, ice-core variations etc., and others based on biological principles like amino acid racemization. However, radiocarbon dating is, by far, the most commonly used science-based dating technique in archaeology.

1.3.2 The rôle of radiocarbon dating in archaeology

Radiocarbon dating has been of principal importance for European archaeology and has resulted in what is known as the 'radiocarbon revolution' (Renfrew 1973). Before the appearance of radiocarbon dating it was considered that the European civilizations originally came from Egypt and Greece, and that all monuments in Europe were 'inspired' by those sophisticated civilizations. Fifty years ago it was believed that Stonehenge or the megalithic tombs of western Europe were the result of the diffusion of ideas from the Near East civilizations. Within this framework, the Egyptian and the Mesopotamian cultures were seen as the earliest and original civilizations in the old world. Renfrew comments that (Renfrew 1973)

"It comes, then, as a shock to learn that all this is wrong. The megalithic chambered tombs of western Europe are now [radiocarbon] dated earlier than the Pyramids (...). In fact Stonehenge, the remarkable and enigmatic structure, can now be claimed as the world's oldest astronomical observatory. The traditional view of prehistory is now contradicted at every point."

Many changes have arisen since European archaeologists experienced such 'shock'. Apart from rebuilding the whole of the old world's chronology, archaeologists needed to change several basic assumptions of their way of thought, all prompted by the evidence of radiocarbon dating.

In the rest of the world, radiocarbon dating might not have created such a 'revolution' but even there it has proved to be one of the most essential techniques for archaeological research. It has been used to help understand the spread of humans into Polynesia (Kirch *et al.* 1989, Kirch *et al.* 1991), and early human settlement in the American continent (Gowlett 1986, Bada *et al.* 1984), to fix two Mayan calendar systems to our modern calendar (Fedik and Taube 1991) and for hundreds of other applications it has brought both small and large changes in world archaeology.

Although radiocarbon dating is employed in a wide variety of problems, in broad terms, it is used in archaeology for one or more of three major purposes.

- (i) Dating objects - an example is the dating of the 'Shroud of Turin' (see Damon *et al.* 1989b) and learning about its authenticity (see Section 4.4).
- (ii) Dating events or contexts with some type of relative chronological ordering - it is becoming more common in archaeology to have information about relative orderings in time of given archaeological events or contexts (eg. A predates B) arising from other types of dating techniques apart from radiocarbon (stratigraphy, for example). Radiocarbon dating is then used to find an absolute chronology for such events or contexts (see, for example, Section 5.3).

- (iii) Dating all other sorts of archaeological events or contexts - it might be the case that archaeologists do not have information from other types of (relative or absolute) dating techniques and radiocarbon is the only practical way to obtain reasonable dates (see Section 5.2).

Before attempting to analyse a set of radiocarbon determinations for its interpretation within a specific archaeological problem, one basic distinction needs to be understood. This is the distinction between dating objects, (i) above, and dating (archaeological) contexts, (ii) and (iii) above. We discuss this problem in the next Section.

1.4 Interpreting radiocarbon determinations

As we have said before, radiocarbon dating is a technique used to estimate the age of some types of objects containing organic matter. In practice, archaeologists find a wide range of objects or materials suitable for radiocarbon dating, from which they select some to be radiocarbon dated. The radiocarbon laboratory then performs the necessary analysis on these and returns to the archaeologists the corresponding radiocarbon determinations. Archaeologists (and statisticians working with them) then have to analyse and interpret the estimates for the age of these objects found on the site under study. But, what was the purpose of dating such objects in the first place? Why estimate the age of a heap of charcoal, a piece of wood or a collection of beans?

Apart from very specific exceptions, radiocarbon determinations are of no interest if viewed as estimates of the age of isolated objects. It is only when archaeologists relate such objects, and the corresponding determinations, to the *context* or *contexts* they are working with that radiocarbon dating proves its full worth and importance. Bowman has explained this point clearly stating that

(Bowman 1990, p. 50)

"The archaeologist is therefore faced with an incomplete and unrepresentative set of data from which a complete whole must be inferred. A process of logic is used to link past events with contexts and features, such as stratigraphic levels and post holes, and to link these with artefacts found within them. If the artefact is organic it can be radiocarbon dated, but it is rare that a date for the artefact *per se* is required; instead it is assumed that the radiocarbon result will also date the event."

How should we use radiocarbon determinations to date 'contexts and features' found on archaeological sites? We believe that the first step to be taken towards solving this question is to recognise the fact that the interpretation of radiocarbon determinations should be *viewed* in relation to what 'contexts and features' are needed to be dated.

As we have pointed out above, radiocarbon determinations are only of use for archaeology when related to the archaeological contexts under study. Therefore it is crucial that the statistical techniques used for the interpretation of radiocarbon determinations include considerations about the relationship between the radiocarbon determination available and the 'contexts and features' of interest. We believe that only then can a correct interpretation of radiocarbon determinations be achieved and thus radiocarbon determinations can be used to date those 'contexts and features'. Furthermore, other dating information might be present in a specific problem. For example, some stratigraphical relationships between contexts might be known. The interpretation of the corresponding radiocarbon determinations ought to consider this and thus the statistical techniques used should allow for all these factors to be included in the analysis. However, apart from some isolated examples, this will involve far more complicated techniques than the calibration of a single determination explained above. Throughout the thesis we will see that the Bayesian framework provides a suitable methodology for developing such statistical techniques.

1.5 Plan of thesis

The thesis will be developed as follows. In Chapter 2 we present a bibliographical review of the most important papers on the statistical analysis of radiocarbon determinations. In this review we try to identify the typical problems of the interpretation of radiocarbon determinations considered in the literature and the inadequacies and limitations of the existing statistical techniques used to tackle them. Based on this we then try to identify the crucial factors that need to be considered for the correct interpretation of radiocarbon determinations.

In Chapter 3 we analyse the radiocarbon calibration problem and develop a novel calibration procedure. We present some simple examples using this new calibration procedure and compare it with other existing techniques. In Chapter 4 we develop a (Bayesian) statistical framework for the interpretation of radiocarbon determinations. This framework tries to overcome the limitations of other approaches reviewed in Chapter 2 by allowing archaeological considerations to be explicitly included in the analysis and interpretation. The calibration procedure developed in Chapter 3 is used and a series of techniques for finding the posterior distributions of interest are presented. In Chapter 5 we then give a variety of examples, showing how our framework is applied in different situations. The typical problems of interpreting radiocarbon determinations reviewed in Chapter 2 are considered and solutions from within our framework are proposed.

In Chapter 6 we analyse the robustness of our framework to the presence of outliers. The problem of outliers in radiocarbon dating is discussed and a novel approach for their identification is proposed. Our framework (developed in Chapter 4) is then extended to allow for the presence of outliers. Two examples are presented where this extended framework is applied and an outlier identification procedure is carried out in each case. In Chapter 7 we present a far more detailed example where our extended statistical framework is used to

interpret a series of radiocarbon determinations arising from an archaeological site in Germany. The archaeological characteristics of the site are studied and, using our extended framework, we develop a statistical model for the analysis of the radiocarbon determinations available. An outlier identification procedure is carried out and the radiocarbon determinations are then interpreted in a way that is consistent with the archaeology of the site. In Chapter 8 we present some concluding thoughts on our work.

Chapter 2

Bibliographical review

2.1 Introduction

In this Chapter we review the relevant publications about the analysis and interpretation of radiocarbon determinations. There are now hundreds of papers relating to the topic of radiocarbon dating, including a journal (*Radiocarbon*) exclusively dedicated to the subject. A large proportion of the publications are devoted to the analysis of the chemical and physical processes involved. We will not study this problem, since the radiocarbon dating process is a highly-specialised scientific area. Instead, we focus attention from the point of view of the users of radiocarbon (principally archaeologists) and on the analysis and interpretation of a set of radiocarbon determinations reported by the laboratories. We consider three main points, namely,

- (a) the reliability of radiocarbon dating,
 - (b) the calibration of radiocarbon determinations
- and
- (c) the interpretation of radiocarbon determinations.

In the next Section we mention briefly some publications that shed light on point (a). This is done so that we may understand the reliability of radiocarbon dating in order that we can develop our work appropriately in the succeeding Chapters. A more critical review is presented in Sections 2.3 (calibration) and 2.4

(interpretation) where we discuss the publications relevant to points (b) and (c). The review of points (b) and (c) represents an important part of our work, where we identify the typical problems of calibration and interpretation of radiocarbon determinations and analyse the limitations of the techniques other authors have used to tackle them. At the end of the Chapter we give a final discussion identifying the problems to be considered later in the thesis.

2.2 Reliability of radiocarbon dating

2.2.1 Introduction

Below we present a brief outline of some interlaboratory studies directed at assessing the reliability of the radiocarbon dating process. We do not intend to give an in-depth analysis of these studies, nor do we present further research on this subject (for a more comprehensive review see Scott *et al.* 1990 and Scott *et al.* 1990b). Rather, we are exclusively interested in this topic from the point of view of the radiocarbon user. From this perspective it is important to establish how reliable the radiocarbon dating technique is and, given this information, to make realistic interpretations of radiocarbon determinations.

Three major interlaboratory studies have been undertaken in the last 15 years, the last one being the most important and extensive. In the next Section we describe all these studies and focus particular attention on the last one (the International Collaborative Study, ICS) discussing its developments and conclusions.

2.2.2 Overview of some interlaboratory studies

The first formal interlaboratory study is reported in Otlet *et al.* (1980). Eight laboratories, all from the UK, finished the study. Five sets of benzene samples

were prepared having ^{14}C levels equivalent to radiocarbon ages of 20,000, 10,000, 5,000, 2,000 and 200% modern. The benzene samples needed little or no pretreatment and thus the only source of variability was related to the counting process. The study concluded that results from the eight laboratories were in agreement and that the standard errors quoted satisfactorily reflected the uncertainties in the process. However, the study recognised its limitations and recommended further interlaboratory comparisons including a wider range of samples and a world-wide laboratory participation.

The first international interlaboratory study was carried out in the early 1980's and the summary of its results are contained in ISG (1982). This project is known as the International Study Group (ISG). 20 radiocarbon laboratories from around the world completed the study. The laboratories were asked to routinely date eight wood samples referred to as *time points* 1-8. A piece of wood of 200 years growth was used to provide all the samples. Eight sets of samples were cut, each set containing samples of the same 10 year tree-ring growth. For each time point the laboratories were therefore measuring the same radiocarbon age and thus the corresponding determinations could be compared.

The determinations obtained at each time point were compared against 'consensus values' (eg. the weighted average of determinations). The variability of the determinations was then assessed and this compared with the standard errors reported by the laboratories (or σ in a determination $y \pm \sigma$). The variability was analysed for individual laboratories and for the study group as a whole.

The ISG observed some 'systematic laboratory bias (...) and a level of variability not entirely explained by the quoted error'. For some time points, differences were observed of up to 700 radiocarbon years. It was concluded that quoted standard errors needed to be multiplied by a factor of between 1.65 to 3.0 to obtain a more realistic standard error. The study group recommended 'further

research of an intercalibrative nature involving different sample types and ages'.

The most extensive interlaboratory study is known as the International Collaborative Study (ICS), see Scott *et al.* (1989), Aitchison *et al.* (1990b) and Scott *et al.* (1990c). 50 laboratories participated in the study over a period of 4 years. Several samples were given to the laboratories to be radiocarbon dated at three different stages. At each stage, the sample pretreatments increased in complexity. In addition, replicated samples were submitted for dating both within and across the laboratories. Therefore, the internal and external consistency of the laboratories could be assessed. To achieve this three main points were considered.

- (i) Internal laboratory variability. That is, how consistent each laboratory is when dating samples of the same radiocarbon age, and to what extent their quoted errors explain the observed variability.
- (ii) Variability due to sample pretreatment and laboratory type. That is, to what extent pretreatments and laboratory type (conventional and AMS dating) influence the dating process.
- (iii) External variability and systematic laboratory biases. That is, the consistency between laboratories when dating samples of the same radiocarbon age, and to what extent their quoted errors explain the observed variability.

With respect to internal variability, the radiocarbon determinations of replicate samples performed by each laboratory were compared and from this the internal consistency of each laboratory was assessed. In relation to the external variability, for each group of samples a consensus value was calculated ('the median') and using this the 'laboratory offset' was measured. Based on this offset systematic biases were measured and the external variability was assessed.

By comparing results from the three stages, it was evident that pretreatments increased variability in the dating process. In addition some differences in the source of variability was observed for the different laboratory types. Of even greater concern was the fact that at each stage, outlying determinations appear with some frequency, with offsets of up to 500 radiocarbon years. In this sense Scott *et al.* (1990c) conclude that, in general, the quoted standard errors do account for internal laboratory variability. However, they also conclude that systematic biases occur between laboratories and that they find (Scott *et al.* 1990c)

"widespread evidence that quoted errors do not adequately describe the variation amongst laboratories (...)"

In the concluding report of the international workshop on intercomparison of radiocarbon laboratories Baxter (1990) states that, of the laboratories participating in the ICS

"Two labs grossly overestimate errors but most labs seriously underestimate their errors by a factors of 2 to 3 times. Only 7 labs from 38 passed all three very basic desirable performance criteria (...)"

To this, Baxter adds that laboratory 'bias, of 50-250 years, is common'. On the whole these findings do not differ greatly from the previous conclusions of the International Study Group (ISG).

Although these conclusions might seem alarming, the same ICS has helped the radiocarbon community to detect some sources of error and improve the overall quality of their analyses. Baxter (1990) comments that

"many labs have already used this study [ICS] to identify their problems, change their procedures and reduce errors"

Furthermore, protocols for quality assurance have been developed to improve the performance of the radiocarbon laboratories (see Switsur 1990, Long and Kalin 1990). In this respect, Scott *et al.* (1990c) state that

"Users of radiocarbon dates may be assured of the continuation of a program for improvement in what is a complex scientific field."

2.2.3 Discussion

In the light of the above discussion and from the point of view of the radiocarbon user, we believe that two major issues should be considered in relation to the reliability of radiocarbon dating. Firstly, due to improved quality control protocols and to the permanent commitment of the radiocarbon laboratories to providing better results, it is reasonable to expect a 'good' reliability from radiocarbon dating (specially for radiocarbon analyses performed after the late 1980's). Secondly, given the complexity of the whole radiocarbon dating process and supported by the evidence obtained from the interlaboratory studies, errors can be expected. Therefore we believe that in interpreting radiocarbon determinations a compromise between credibility and caution should be undertaken with respect to the reliability of the radiocarbon dating technique. This compromise can be summarised by two major points.

- (i) The radiocarbon user should expect the best quality control and performance from the radiocarbon laboratories. Thus a radiocarbon determination $y \pm \sigma$ returned by a laboratory may be assumed to be a reliable estimate of the radiocarbon age for the sample analysed. Furthermore, σ may be considered to be a realistic measure of the error in the dating process.

- (ii) Point (i) should be treated with caution since mistakes in the dating process cannot be ruled out. This implies that the presence of erroneous (outlying) determinations, with offsets ('shifts' in the radiocarbon age) of any magnitude, cannot be dismissed when interpreting radiocarbon determinations.

In this thesis, point (i) is assumed when developing our calibration method and the statistical framework for the interpretation of radiocarbon determinations (in Chapters 3 and 4, and 5 respectively). This is, of course, in relation to our basic model which assumes that, given a radiocarbon determination $y \pm \sigma$, $y \sim N(\mu, \sigma^2)$, where μ is the 'true' radiocarbon age for the sample dated and σ is *known*. Thus no further error or correction is introduced and σ is assumed to represent a realistic measure for the error in the determination. However, in Chapter 6 point (ii) is considered and the robustness of our framework to the presence of outliers is analysed. A novel approach for the problem of outliers in radiocarbon dating is proposed and our statistical framework is extended to allow for the presence of outliers. This extended framework then takes into account in a more realistic way the reliability of the radiocarbon dating technique discussed here.

2.3 Calibration

2.3.1 Introduction

By radiocarbon dating dendrochronologically dated samples, the accuracy of radiocarbon determinations was assessed. At the early stages of radiocarbon dating some imprecisions became apparent but it was difficult to distinguish systematic errors from the standard errors reported for the determinations. Two basic assumptions Libby made when developing radiocarbon dating were that,

(i) the atmospheric ^{14}C level is uniform around the globe

and

(ii) the atmospheric ^{14}C level has remained constant through time.

While assumption (i) above has been found to be reasonable on the basis of experimental data (although very recently some doubts about this have surfaced, see Pearson and Stuiver 1993), assumption (ii) was challenged. De Vries (1958) demonstrated that there have been some significant changes in the atmospheric ^{14}C , at least over the last 400 years. A more detailed study by Willis *et al.* (1960) covering the last 1300 years came to the same conclusion. Soon theoretical arguments appeared explaining the genesis of such variations (related to geomagnetism and solar activity) and by the mid 1960's it was generally accepted that assumption (ii) was incorrect.

Thus radiocarbon determinations reported in radiocarbon years had to be calibrated onto the calendar scale and therefore measured in calendar years in order to be of use in archaeology. The elements of a radiocarbon calibration data set are divided in three parts, namely

(i) a calendar year t (obtained by the dendrochronological date of the tree-ring sample),

(ii) a radiocarbon year x

and

(iii) the associated standard error σ for x .

Therefore a calibration data set consists of

$$\{[t_1, x_1 \pm \sigma_1], [t_2, x_2 \pm \sigma_2], \dots, [t_n, x_n \pm \sigma_n]\}.$$

How can we use the $[t_k, x_k \pm \sigma_k]$'s to calibrate a radiocarbon determination?

Firstly, we assume the existence of a function $r(\cdot)$ such that, for any given calendar year θ , $r(\theta)$ is the corresponding true radiocarbon year. That is, $r(\theta)$ is the calibration curve. Thus $x_k = r(t_k) + e_k$ and, assuming normality,

$$x_k \sim N(r(t_k), \sigma_k^2).$$

Secondly, from the calibration data we need to approximate $r(\theta)$ and then develop a methodology to transform a given radiocarbon determination $y \pm \sigma$ onto the calendar scale. Broadly speaking this represents the problem of radiocarbon calibration. Below we discuss the calibration data set to be used in the thesis and review the relevant publications on radiocarbon calibration.

2.3.2 The radiocarbon calibration data

Since the late 1960's much effort has been invested by the radiocarbon community to radiocarbon date dendrochronologically dated samples and thus obtain estimates of the radiocarbon age of wood of known calendar age (calibration data). At the 'Twelfth Nobel symposium' held at Uppsala University in 1969 three calibration data sets were presented, and are described in Damon (1970), Ralph and Michael (1970) and Suess (1970). Of these three data sets the Suess data is the most extensive, ranging from 5200 BC to the present.

In the early 1970's several questions about the radiocarbon calibration data were still unresolved. The most basic concern was related to the global

applicability of the results. That is, would the ^{14}C level found in a series of rings taken from a specific tree be a reliable estimate of the worldwide atmospheric ^{14}C level at the time indicated by those rings? By comparing the ^{14}C level of modern tree-ring samples from around the globe, Ralph *et al.* (1973) conclude that the ^{14}C levels are in good agreement and, 'on the average', independent of the origin of the tree-ring samples. Therefore, global applicability of the calibration data can be assumed. By combining the calibration data sets mentioned above Ralph *et al.* (1973) construct a calibration data set spanning the period 1849 AD to 4769 BC. This calibration data set, together with that of Suess, were frequently used in the 1970's for the calibration of radiocarbon determinations.

In order to accurately calibrate radiocarbon determinations we need calibration data to be as precise and reliable as possible. Using larger tree-ring samples and following an international agreement on the standards to be observed for radiocarbon dating, a huge effort has been made to create a high-precision calibration data set, with standard deviations as low as 10 to 20 years.

The systematic high-precision radiocarbon dating of tree-ring samples began at Seattle in 1973 (M. Stuiver) and at Belfast in 1975 (G. W. Pearson). The results of these two research groups is contained in Stuiver and Pearson (1986), Pearson and Stuiver (1986) and Pearson *et al.* (1986), and constitutes what the radiocarbon community refers to as the 'Internationally agreed high-precision calibration curve'. We will refer to this data as the 'high-precision calibration data'. At present the international radiocarbon community has agreed that this is the data which should be used for radiocarbon calibration, and therefore we will use this data in this thesis. (Very recently other calibration data sets have appeared although the whole of the international radiocarbon community has not to agree on their usage, see Pearson and Stuiver 1993.)

2.3.3 Early approaches for calibration

One of the first published estimates of the radiocarbon calibration curve, $r(\theta)$, is given in Suess (1970) where 'The curves are drawn by hand'. The calibration procedure recommended basically involves calculating $r^{-1}(y-\sigma, y+\sigma)$ where $y \pm \sigma$ is the determination to be calibrated. That is, the inverse image of the one-sigma region $(y-\sigma, y+\sigma)$ is computed over the calibration curve (we discuss this calibration procedure in greater detail below). In relation to this Suess (1970) explains that

"Because of the peculiar windings of the calibration curve, one particular radiocarbon content may indicate several [calendar] dates (...). In general, these several dates will lie within the limits of experimental error of the particular radiocarbon measurements and then, in such cases, it will only be possible to establish the time interval during which the sample had originated. For certain periods of time, this time interval will be quite large."

Indeed, due to a combination of long-term and short-term components in the variations of the atmospheric ^{14}C , 'kinks' or 'wiggles' are found in the calibration curve. It took a long time for the radiocarbon community to accept that the variations observed in the calibration data are a result of the complicated nature of $r(\theta)$ and not a result of misleading or erroneous data. It is now generally accepted that such wiggles are an inherent feature of $r(\theta)$ and should not be totally smoothed by an estimation procedure. The result is that the estimated calibration curve will be non-monotonic and 'wiggly'.

Suess's estimate was one of the first given for the calibration curve. Although these estimation and calibration methods are quite informal, undoubtedly they provide a starting point for the understanding of the radiocarbon calibration problem.

A more systematic approach was developed by Ralph *et al.* (1973). They combine the different calibration data available at that time and create a single calibration data set. After testing some smoothing procedures to approximate $r(\theta)$ they conclude that (Ralph *et al.* 1973)

"a 9-cell floating average centered on its mid-point seems to be the best choice since it resulted in a relatively smooth curve, but it did also preserve the major deviations and most of the minor ones expressed in the raw data (...)"

Thus it seems that the major preoccupation when estimating $r(\theta)$ is,

(i) to have a smooth curve

and

(ii) to follow the 'kinks' in the data (ie. the estimated curve should not be too smooth).

The procedure Ralph *et al.* (1973) used for calibration is clearly informal based on common-sense and intuition. However, they do give an extensive table of radiocarbon years in decades and their corresponding calendar years. Since $r(\theta)$ is non-monotonic, a radiocarbon year y can correspond to one or several calendar years. Ralph *et al.* (1973) divide this into three cases.

"The majority of the corrections for radiocarbon dates (...) are found to be single crossings. If the radiocarbon date follows the curve closely for a distance (usually a relatively short one) we designate this distance as a span. In those cases where a radiocarbon date crosses the curve two or more times, we must consider an overall range (...) Of course, it is possible for a range to have a span and/or a crossing within the range (...) An example of the use of tables for a single crossing is as follows: The radiocarbon date of A.D. 1750 crosses the curve at the dendrochronologically determined date of A.D. 1650. Thus the correction for this date is 100 ± 10 years (...) The ± 10 years expresses the statistical uncertainty of the [calibration curve] and must be added to the standard statistical error (...) of the radiocarbon

date."

The basic problem with this procedure is that Ralph *et al.* continue with the concept of a date with an associated standard error, even after calibration. Given the normal model used, a (uncalibrated) radiocarbon determination can be satisfactorily summarised in terms of a mean and a standard error ($y \pm \sigma$). However the corresponding distribution in the calendar scale of a radiocarbon determination is complicated, often multimodal, and sometimes not very smooth. Therefore it is not necessarily true that a calibrated date can be satisfactorily summarised using a central date and an error term. At present this procedure is not commonly used, and in general the radiocarbon community does not consider it valid. This procedure is not used in the thesis.

2.3.4 Two major contributors

A critical review of calibration can be found in Renfrew and Clark (1974) where several implications of the radiocarbon calibration process are discussed. They note that (Renfrew and Clark 1974)

"Radiocarbon determinations of dendrochronologically dated samples cannot simply be applied directly and without further thought to yield 'calibrated dates' "

They define the calibration function $r(\theta)$ and the 'inverse calibration function' ($r^{-1}(\theta)$) but correctly argue that such a 'function' does not exist due to the non-monotonicity of $r(\theta)$. From that it is clear that we need to approximate $r(\theta)$ and not its 'inverse'. Although they attempt to approximate $r(\theta)$ using some smoothing procedures they explain that (Renfrew and Clark 1974)

"Calibration of radiocarbon dates requires only the existence of some smooth curve, which is simply a good fit to the data (...)"

They study the calibration procedure in the following way. Firstly they consider the case in which $r(\theta)$ is exactly known. This might be the case when we ignore the standard errors in the calibration data (the σ_k 's defined earlier). Then to calibrate a radiocarbon determination $y \pm \sigma$ they consider the set

$$C = r^{-1}(y - \gamma\sigma, y + \gamma\sigma).$$

That is, the inverse image over $r(\theta)$ of the confidence set $(y - \gamma\sigma, y + \gamma\sigma)$, where the confidence α of the set is given by the value of γ . The resulting C will then be a 100 $\alpha\%$ 'calibrated confidence set'. Usually C will consist of several unconnected intervals. We call this the 'confidence intervals calibration method'.

A problem we note with this procedure is that, since the radiocarbon age is not considered to be uniformly distributed over $(y - \gamma\sigma, y + \gamma\sigma)$ and $r(\theta)$ is not linear, the resulting distribution for the calendar age will not be uniform over C . The set C is a union of open intervals (since $(y - \gamma\sigma, y + \gamma\sigma)$ is open and $r(\theta)$ continuous) but the probability of each interval is not necessarily proportional to its length. This could easily lead to erroneous interpretations and, indeed, potentially important information is being lost. As seen in Chapter 1, it is not difficult to calculate the whole of the resulting distribution in the calendar scale when we suppose $r(\theta)$ to be known (in that case we use the piece-wise linear calibration curve $\mu(\theta)$) and we see no reason why a confidence interval like C should be preferred.

Secondly, Renfrew and Clark consider the case in which $r(\theta)$ is known but only approximately. They then suppose that $r(\theta) \in (\varepsilon_1(\theta), \varepsilon_2(\theta))$, where the functions $\varepsilon_1, \varepsilon_2$ represent the 'confidence band' for the calibration curve. They propose a procedure for calibration which consists in considering

$$C = \varepsilon_1^{-1}(y - \gamma\sigma, y + \gamma\sigma) \cup \varepsilon_2^{-1}(y - \gamma\sigma, y + \gamma\sigma).$$

That is, the inverse image of the confidence interval $(y - \gamma\sigma, y + \gamma\sigma)$ over the 'confidence band' for $r(\theta)$. The principal problem with this procedure is that C does not represent a confidence set of the same confidence level as $(y - \gamma\sigma, y + \gamma\sigma)$. C is bigger than necessary. They sketch a procedure to find 'exact' confidence sets C , but it is not developed in detail.

A more formal approach can be found in Clark (1979). Here Clark studies the calibration data sets available and states a general model for calibration. He casts doubt about the general applicability of the calibration data and in his statistical model explicitly introduces distinctions between different calibration data sets. This is done in the following way. Let $[t_k, (x_{k,j} \pm \sigma_{k,j})]$ be the calibration observation measured at laboratory j for the calendar year t_k , then the model states that

$$x_{k,j} = r(t_k) + H_j + f_k + e_{k,j}.$$

The terms H_j 's represent 'systematic errors between laboratories', the f_k 's 'the intrinsic variability of contemporaneous samples due to local changes in the ^{14}C levels' and $e_{k,j}$ 'the net errors of measurements' arising from the $\sigma_{k,j}$'s.

Clark then explains that, due to evidence about the uniformity of the levels of the atmospheric ^{14}C around the globe, and due to the improved quality assurance in the radiocarbon process, the following assumptions can be made.

- (i) There are no systematic errors between laboratories ($H_k = 0$).
- (ii) The distribution of ^{14}C in the earth's atmosphere is uniform and varies uniformly around the world ($f_k = 0$).

According to the above assumptions Clark states his simplified model

$$x_k = r(t_k) + e_k$$

where $[t_k, (x_k \pm \sigma_k)]$ are the combined calibration observations for all laboratories and $e_k \sim N(0, \sigma_k^2)$. One of the main features of this model for calibration is that it does include the standard errors reported in the calibration data (the σ_k 's).

Clark then uses a 'convolution-smoothing' (CS-) estimator to approximate $r(\theta)$. He explains that

"(...) because of the computational difficulties in fitting cubic splines to over 700 data-points, we used (...) a simple but similar method in which the estimate of (...) $[r(\theta)]$ was chosen by cross-validation from a class of convolution-smoothed first order interpolating splines (...)"

The resulting calibration curve is 'almost linear with very few wiggles'. Using the 'confidence intervals calibration method' explained above Clark (1979) develops a procedure to calculate 'exact' confidence regions. However, this procedure is difficult to implement and 'unwieldy to be useful in practice'. Clark then calculates a 'somewhat conservative' confidence region that for the case of the CS-estimator used for $r(\theta)$ is

$$\{\theta: y-d < f(\theta) < y+d\}$$

where $f(\theta)$ is the CS-estimator of the calibration curve ($r(\theta)$) and d is a positive number dependent upon the error related to the estimation of $f(\theta)$, the error of the determination (σ) and upon the chosen confidence level. These confidence regions are similar to those suggested before by Renfrew and Clark (1974) but with a correct assessment of d .

In addition to the inadequacies of this 'confidence intervals calibration method' alluded to earlier, the extension of this approach to more complicated

scenarios, where one has a set of radiocarbon determinations and perhaps prior information relating them, is not clear. A particular scenario of this kind is considered by Clark (1979) in which, for a set of radiocarbon determinations $\{y_1 \pm \sigma_1, y_2 \pm \sigma_2, \dots, y_m \pm \sigma_m\}$, associated with the calendar years $\theta_1, \theta_2, \dots, \theta_m$, we have $\theta_j = \alpha + l_j$, where the l_j 's are known constants and α is unknown. That is, we *know* the differences of the associated calendar years for any two determinations $y_j, y_{j'}$ ($\theta_j - \theta_{j'}$), but we do not know their absolute position in the calendar scale. The problem then is to estimate α (and consequently have estimates for all the θ_j 's) using the radiocarbon determinations.

The above problem is known in archaeology as dating a 'floating chronology'. Cases like this normally arise when we have samples from tree-rings and the differences $\theta_j - \theta_{j'}$ are obtained by counting the rings separating each sample. Clark sketches four methods for estimating α of which, the first three basically use the same technique as for simple calibration. For method four he explains that

"In the case of a floating chronology from an archaeological site, there may be considerable prior information concerning α , expressible as a prior density. If our prior information regarding (...) $r(\theta)$ can also be expressed in terms of prior densities, one can use standard Bayesian methods to give a posterior density for α "

However, he only suggests the idea and does not develop a technique based on Bayesian methods to estimate α . (In Section 5.3 we return to this problem of dating floating chronologies.)

2.3.5 Recent developments

Pearson and Stuiver (1986) use the high-precision calibration data to develop a calibration technique based on a piece-wise linear calibration curve. They recommend transforming the interval $(y - \sigma, y + \sigma)$ onto the calendar scale

(basically using the set $r^{-1}(y-\sigma, y+\sigma)$) but no means are provided for assigning probability to the resulting region on the calendar scale. However, they seem to recognise that such procedure was only provisional since (Pearson and Stuiver 1986)

"The non-linear transform of a Gaussian standard deviation around a ^{14}C age into calendar AD/BC (BP) ages leads to very complex probability distributions that can only be calculated with the aid of computers. We are currently developing suitable programs for these probability calculations (...)."

The 'computer programs' are now widely available and the software is known as 'CALIB' (Stuiver and Reimer 1986, 1993). At present, it appears that CALIB is the most commonly used computer program for the calibration of radiocarbon determinations.

Besides CALIB, other authors have developed computer programs for the calibration of radiocarbon determinations using the high-precision calibration data. Aitchison *et al.* (1989) present a comparative study of the eight most well known of such programs. In the programs studied there is no agreed method for calibration. Furthermore, the presentation (graphics, layouts etc.) and other details vary greatly from program to program. However, for the calibration of a single radiocarbon determination $y \pm \sigma$ there are only two basic methods used.

(A) Calculating $C = r^{-1}(y - \gamma\sigma, y + \gamma\sigma)$, (the 'calibrated confidence intervals method').

(B) Calculating $f(\theta) = K \exp\left\{-\frac{(y-r(\theta))^2}{2\sigma^2}\right\}$ (the 'probability distribution method').

All the programs use variations of these two methods. Aitchison *et al.* (1990) follow method (A). Leese (1988), Otlet ('personal communication'), Pazdur and Michczynska (1989), van der Plicht *et al.* (1990), Robinson (1986) and Weninger (1986) follow method (B). Stuiver and Reimer (1986) (CALIB) offers the option of using either. The variations include the following.

- (i) The choice of the calibration curve $r(\theta)$.
- (ii) For method (A), the choice of γ and the graphical layout for the confidence intervals in C .
- (iii) For method (B), the choice of K and the presentation of $f(\theta)$.

In all but one of the programs the choice for $r(\theta)$ is the piece-wise linear calibration curve $\mu(\theta)$. Van der Plicht *et al.* (1990) uses a cubic spline fitting approximation, but the resulting calibrated distributions are almost identical to those obtained using $\mu(\theta)$.

As mentioned earlier, Clark (1979) studies method (A) and gives exact and 'conservative' values for γ to obtain calibrated confidence intervals of correct size. Thus if method (A) is preferred, Clark's technique should be used. However, we have explained the inadequacies of this calibration method and we prefer not to use it.

Method (B) is a particular case of the calibration method we use in this thesis, with K defined so that $f(\theta)$ is normalised to one. Some authors do not insist on normalising $f(\theta)$ (Weninger 1986) and discuss alternative choices for K . In fact, Stuiver and Reimer (1993) allow the option of choosing K so that the maximum of $f(\theta)$ is 1. We see no sensible reason for not normalising $f(\theta)$ to one,

particularly as it will be interpreted as a probability density. In the programs that use method (B), a combination of histograms and highest posterior density (HPD) regions for different probabilities and quantiles are given. However, we do not intend to present a detailed comparison of the graphical methods used by each of the programs for the presentation of histograms, HPD regions or the like.

With respect to the standard errors reported in the calibration data (the σ_k 's), various approaches are followed. Pazdur and Michczynska (1989), Robinson (1986) and van der Plicht *et al.* (1990) decided to neglect them. Otlet, Leese (1988) and Weninger (1986) calculate $s^2 = \sigma^2 + \sigma_c^2$, where σ_c estimates the standard deviation in the curve based on the σ_k 's and is commonly fixed at 10 or 20 years. From that they calibrate $y \pm s$. A third method, used by Aitchison *et al.* (1990) and Stuiver and Reimer (1986) is to give a 'confidence band' for $r(\theta)$ centred at the piece-wise linear calibration curve. That is, $r(\theta) \in (\mu(\theta) + \sigma(\theta), \mu(\theta) - \sigma(\theta))$ and $\sigma(\theta)$ is a linear interpolation of σ_k and σ_{k-1} , where $t_k > \theta \geq t_{k-1}$. From this Renfrew and Clark (1974) calibrated confidence intervals method is used (discussed above).

None of these methods for including the standard errors σ_k 's in the calibration process are thoroughly justified by their authors. Under certain conditions each of these methods can be justified. However, the estimates obtained for the standard deviation in the calibration curve can only be seen as informal approximations. In this thesis we prefer to develop a stochastic model for $r(\theta)$ using observational data for the atmospheric ^{14}C levels through time. From that, we formally estimate its standard deviation based upon the σ_k 's. This represents the basis for our novel approach to calibration, which includes the standard deviations σ_k 's (this will be developed in Chapter 3).

2.3.6 Discussion

We believe that before deciding on the specific graphical presentation and layout for any calibration results, a satisfactory method for calibration must be developed. Based upon the Bayesian framework one can see that the correct method for calibration, using the piece-wise linear calibration curve and neglecting the errors in the calibration data, is the one given in Chapter 1. For the specific case of vague prior information about θ , method (B) above will give the same results as the Bayesian method. From this a wide range of possibilities can be used to present and summarise $f(\theta)$, from a simple histogram to complicated graphical layouts to plot HPD regions or quantiles of $f(\theta)$. Under certain circumstances, however, the standard deviation of the calibration curve should be considered. In Chapter 3 we discuss this issue and develop the calibration procedure to be used in the thesis.

In summary, it can be seen from the discussion in the previous Sections that there is no single established method with which to calibrate radiocarbon determinations. On the contrary, each researcher (or group of researchers) tends to have their own preferred calibration method and to present their results differently. An even bigger problem occurs when we need to introduce archaeological information into the analysis. Only two of the eight studied programs claim to have the facility to do so. The first is a basic implementation of Aitchison *et al.* (1991), and we discuss their approach in Section 2.4.3. The second is an implementation of the ideas presented by Weninger (1986), which we discuss in Section 2.4.4.

2.4 Interpretation of radiocarbon determinations

2.4.1 Introduction

We now discuss some of the relevant publications concerning the interpretation of radiocarbon determinations. In doing so we consider a wide variety of problems and study several different statistical techniques. We attempt to identify typical problems in the analysis and interpretation of radiocarbon determinations, and the techniques others have previously used to tackle them. Throughout the review we focus attention on the crucial factors that must be considered in any interpretation problem. We highlight both the strengths and weaknesses of the existing techniques.

This Section of the bibliographical review represents a central part of the thesis. It is here that we try to identify the underlying problems in the analysis and interpretation of radiocarbon determinations. The majority of the papers considered below are concerned with specific interpretation problems. We will, however, try to avoid discussing details relevant only to particular examples and concentrate on the global characteristics of the problems. This will then provide a basis for our own general approach to be developed in later Chapters.

2.4.2 Early work

Libby (1954), in one of the first published list of radiocarbon determinations, briefly discusses the interpretation of a set of determinations. The archaeological dating problem studied by Libby is related to the ascension of 'Hammurabi of Babylon'. The Babylonian calendar gives dates for the ascension of some of Hammurabi's predecessors over a period of at least 350 years. However, this calendar provides only relative dates since its relationship to our calendar is not completely known. A charred beam from the remains of a house thought to be of

the reign of a predecessor of Hammurabi was divided to produce three samples, each of which was radiocarbon dated. By fixing the date of the beam using radiocarbon, it was possible to relate the ancient Babylonian calendar to the modern calendar and hence date the ascension of Hammurabi of Babylon.

To combine the three determinations obtained and give an estimate for the age of the charcoal beam, Libby (1954) proposed using a weighted average 'using the inverse square of the counting errors as weighting factors'. Thus the weighted average is $\sigma^2(y_1\sigma_1^{-2} + y_2\sigma_2^{-2} + y_3\sigma_3^{-2})$ where $\sigma^{-2} = \sigma_1^{-2} + \sigma_2^{-2} + \sigma_3^{-2}$. However, later he writes

"It is probably better, however, to take the arithmetical average since there are undoubtedly other errors than the counting errors."

To obtain a final date, Libby takes the arithmetic mean $y = (y_1 + y_2 + y_3)/3$ although the variance he associates with it is σ^2 as calculated above. Assuming that the age for the sample is normally distributed with mean y and variance σ^2 , he estimates that with a 95% probability the organic materials in the charred beam died between 2205 and 1887 BC.

One obvious problem with Libby's (1954) procedure is that it does not include calibration. This, of course, can be understood since at that time it was not suspected that radiocarbon determinations needed to be calibrated to obtain calendar dates. On the other hand, a very important issue is how and when to average a set of determinations to give a single 'date'. Given the standard Gaussian model for radiocarbon determinations it can be seen that, if we are to average the determinations, a weighted average (or pooled mean) should be preferred over an arithmetic average, in contrast to what Libby did. Of even greater concern is when this should be done. We now discuss this problem in some depth.

One point to be appreciated before attempting a statistical analysis of a set of radiocarbon determinations is the need to distinguish between two basic cases,

Case I, when we have a series of radiocarbon determinations taken from the same object (as in Libby's case above)

and

Case II, when we have a series of radiocarbon determinations taken from different objects.

In this thesis, to distinguish between Case I and Case II, we will call Case I the case representing a set of *replications*, with all other sets of radiocarbon determinations belonging to Case II, unless stated otherwise. There is a crucial difference in the statistical techniques to be used for these two cases. The above distinction was first clearly identified by Ward and Wilson (1978). Previous to that, Spaulding (1958), Polach and Golson (1966) and Leach (1972) considered the statistical analysis of various sets of radiocarbon determinations but do not make this distinction.

Another early work is Law (1975) who performs an analysis of a set of radiocarbon determinations from a site in New Zealand. Law does not distinguish between Case I and Case II, and does not clearly establish any specific statistical model. However, the section 'Association of the sample with the event [trying to be] dated' is of some interest. He explains that contextual information from the site must be taken into account to correctly associate the samples with the events under study. Through this association, it is hoped that the context or events considered could be correctly dated. Unfortunately, Law (1975) was unable to include contextual information in his statistical analysis.

An interesting paper concerning the averaging of radiocarbon determinations is Long and Rippeteau (1974). They propose the usage of the 'pooled mean'

$$y_p = \sigma^2 \sum_{j=1}^m \frac{y_j}{\sigma_j^2} \quad \text{with} \quad \sigma^{-2} = \sum_{j=1}^m \frac{1}{\sigma_j^2} \quad (2.1)$$

and make use of a 'Chauvet's rejection criterion' (Chauvet 1863). This criterion consists of rejecting any observation for which $|y_j - y_p| > \gamma\sigma$ where

$$P[|Z| > \gamma] = 1/(2m)$$

and $Z \sim N(0,1)$. This appears to be an *ad hoc* criterion that has several controversial components, for example, the choice of the threshold $1/2m$ or the dependence of y_j and y_p . The gross inadequacies of this criterion are discussed by Barnett and Lewis (1984), Renfrew and Clark (1974) as well as by Ward and Wilson (1978).

We believe that Long and Rippeteau's discussion on 'When averaging is appropriate' in which they urge us 'to understand non-statistical discrimination in selecting the dates to be averaged' is extremely important. This, albeit implicitly, raises the issue of the archaeological context on which we are working, and how it relates to the statistical techniques we are attempting to use. That is, we believe, one of the principal issues to be taken into account in the interpretation of radiocarbon determinations. In fact, to their credit, Long and Rippeteau (1974) distinguish correctly between Case I 'replicate runs on identical sample material' and Case II. Discussing whether or not to average radiocarbon determinations in Case II problems, they note that this 'will involve judgements based upon archaeological (...) knowledge'. We believe this to be a highly relevant comment since only then can a set of determinations be averaged in a meaningful way. We continue to study this problem in Section 4.4.

2.4.3 Ward and Wilson (1978)

Probably the most influential and widely quoted work on the statistical analysis of sets of radiocarbon determinations is that of Ward and Wilson (1978). The techniques they propose are now routinely used by archaeologists and radiocarbon laboratories. The paper is primarily concerned with how and when it is reasonable to average a set of determinations and obtain a single date.

Ward and Wilson make a large critique of the previous publications up to that time, clearly state the difference between Case I and Case II problems, and propose a significance test for each case. Unfortunately, they did not have the advantage of the availability of a high-precision calibration curve and, as a result, any critique of their paper should be put in perspective. In a later paper, Wilson and Ward (1981), they discuss the problem of outliers but a review of this paper is deferred to Chapter 6, which is specifically concerned with this subject.

The approach Ward and Wilson (1978) follow is based on classical hypothesis testing. Given a Case I set of radiocarbon determinations (replications) $\{y_1 \pm \sigma_1, y_2 \pm \sigma_2, \dots, y_m \pm \sigma_m\}$ they consider the model

$$Y_j = \mu + \varepsilon_j$$

where μ is the 'true' radiocarbon age for all determinations and

$$\varepsilon_j \sim N(0, \sigma_j^2).$$

'To test the hypothesis that the series of determinations are consistent (ie. all have effectively the same [radiocarbon] age)' (they do not state an alternative hypothesis) they propose using the test statistic

$$T = \sum_{j=1}^m \frac{(y_p - y_j)^2}{\sigma_j^2}, \quad (2.2)$$

where y_p is the pooled mean from Equation 2.1. Given the above model, T has a chi-square distribution with $m-1$ degrees of freedom. They then write

"If the determinations are judged not to be significantly different [using statistic T] then they can be combined, the pooled mean being $[y_p]$ (...)"

and the variance of this pooled mean being σ^2 as given in Equation 2.1. They never explain, though, why we should use such a pooled mean to combine the determinations. (In Chapter 4 it will be proved that, under certain conditions, for Case I problems, y_p is a sufficient statistic and is the maximum likelihood estimator for μ .)

The above statistical test is currently widely used and is given in basic texts on radiocarbon dating (see Bowman 1990 p. 58, Gillespie 1986 p. 30, Aitken 1990 p. 95).

Ward and Wilson propose a different model for Case II sets of radiocarbon determinations, which cannot be considered to be replicates. In Case I the model supposes a *unique* radiocarbon year μ common to all Y_j 's, whereas in Case II each Y_j has its own radiocarbon mean μ_j , since the determinations do not arise from the same object. The proposed model is

$$Y_j = \mu_j + \varepsilon_j + f_j + g_j \quad (2.3)$$

where

$$\varepsilon_j \sim N(0, \sigma_j^2)$$

and

$$f_j \sim N(0, \sigma_f^2) \text{ and } g_j \sim N(0, \sigma_g^2).$$

Ward and Wilson do not make use of any calibration data set and thus no explicit

calibration procedure is used by them. However, in this latter model the error terms f_j and g_j are introduced to account for the 'error factor in the calibration curve' and for the 'sunspot effect', respectively. Based on informal considerations, they then estimate that $\sigma_f = 50$ and $\sigma_g = 70$ (taking ε_j , f_j and g_j to be independent).

They go on to consider the following hypothesis test for Case II

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_m$$

$$H_1 : \text{not } H_0, \quad (2.4)$$

to test the hypothesis that all determinations belong to the same radiocarbon year. Then they propose the test statistic T' similar to T (in Equation 2.2) but using $s_j^2 = \sigma_j^2 + \sigma_f^2 + \sigma_g^2$ instead of σ_j^2 . If H_0 is not rejected at some significance level α (that is, $T < Q_{\alpha, m-1}$, where $Q_{\alpha, m-1}$ is the upper 100 $\alpha\%$ quantile of a chi-square distribution with $m-1$ degrees of freedom), then the determinations may be combined using the pooled mean y_p .

Of course, H_0 does not tell us what happens on the calendar scale since the μ_j 's refer to the radiocarbon ages of the sample. Since the calibration curve is non-monotonic, there is not a one-to-one relationship between radiocarbon and calendar years. Therefore, even if we knew that $\mu_1 = \mu_2 = \dots = \mu_m$, the dated objects may have very different calendar ages (θ_j 's), with differences as large as 200 years, see Figure 2.1. Given the availability of the high-precision calibration data, we believe that any statistical analysis should now use a model based on calibration (similar to the one presented in Equation 1.4). Ward and Wilson's approach represents only an approximation relevant only before the calibration data was available. In fact, in relation to the problem considered, it would be necessary to test the hypothesis $H_0' : \theta_1 = \theta_2 = \dots = \theta_n$ against $H_1' : \text{not } H_0'$. That is, the determinations are associated with the same calendar year, rather than

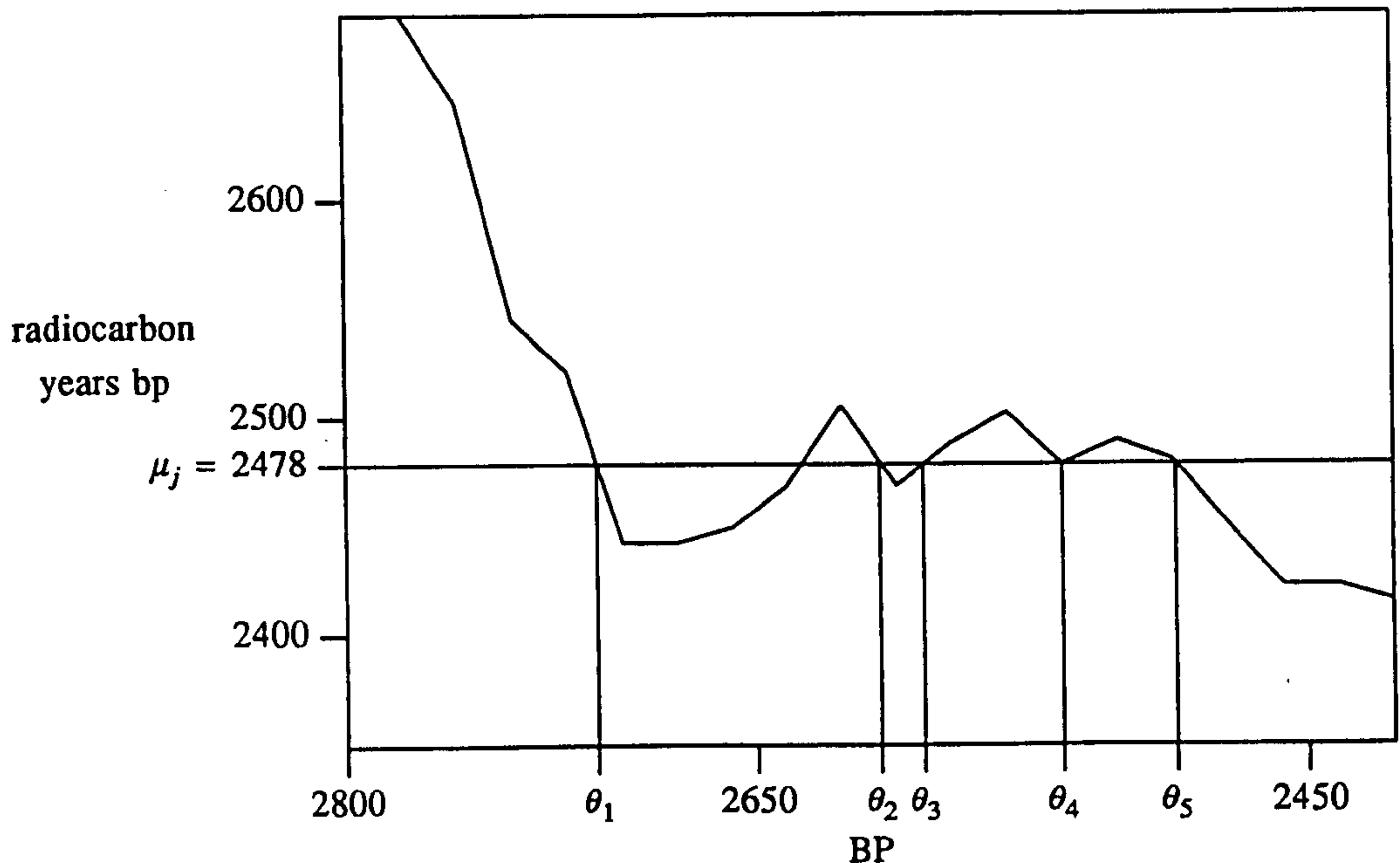


Figure 2.1

The calendar years

$$\theta_1 = 2708, \theta_2 = 2605, \theta_3 = 2589, \theta_4 = 2539, \theta_5 = 2498$$

spanning more than 200 years all have the same radiocarbon age (2478 bp).

testing that they all have the same radiocarbon age, as it is the case above.

Furthermore, despite knowing that the objects have similar calendar ages, we still have the question of whether, archaeologically speaking, it makes any sense to combine the determinations. This problem was mentioned by Ward and Wilson (1978)

"If the estimates of the real dates are judged not to be significantly different [using their T' test] *and*, if from archaeological considerations, it is deemed appropriate, then the radiocarbon determinations can be combined."

and has been addressed by other authors (for example Bowman 1990, p. 60). The sad truth is that the T' -test has been misused (see, for example, Pazdur and Krzaniowski 1991, Saville *et al.* 1987, Nydal 1989, Hassan and Robinson 1987, Aitken 1990 and Stuiver and Reimer 1993), despite the warnings of Ward and Wilson and

other authors. Moreover, there is a tendency to use the uncorrected T -test for hypothesis H_0 (in Equation 2.4) when the problem is obviously of Case II. Then, if H_0 is not rejected, the determinations are combined into a pooled mean which is then calibrated. To investigate the inadequacies of this procedure we consider a set of m calendar years $\theta_1, \theta_2, \dots, \theta_m$ evenly spaced within a period of time $(\beta, \beta + l)$. We then simulate y_j with distribution

$$y_j \mid \theta_j \sim N(\mu(\theta_j), \sigma^2)$$

for some fixed σ^2 , and calculate the corresponding T . We repeat this process n times and record the number of times H_0 would be rejected at significance level α (that is $T > Q_{\alpha, m-1}$, where $Q_{\alpha, m-1}$ is the upper $100\alpha\%$ quantile of a chi-square distribution with $m-1$ degrees of freedom). In Figures 2.2 and 2.3 we find plots of the percentage of rejections for different values of β, l and σ^2 with $m = 5$ and $n=5000$ and $100\alpha = 5\%$.

The peculiar behaviour of the T -test in these circumstances is due to the shape of the calibration curve. Note that this behaviour depends not only on σ^2 but also on the actual position of the period of time $(\beta, \beta + l)$. From Figure 2.2 we see that the procedure is likely to reject H_0 for $l = 150$, $\sigma = 40$ and $\beta = 4050$ BP (more than 50% of rejections), whereas for $\beta = 4550$ BP it is considerably less likely (less than 20% of rejections), as seen in Figure 2.3. This is an unacceptable characteristic of the T -test when wrongly applied to Case II samples.

As part of our critique of Ward and Wilson (1978), we make one general remark that is linked to our general disagreement with the foundations of classical statistics. Long and Rippeteau (1974) in their discussion about the significance (classical) test of hypotheses H_0 , in Equation 2.4, say

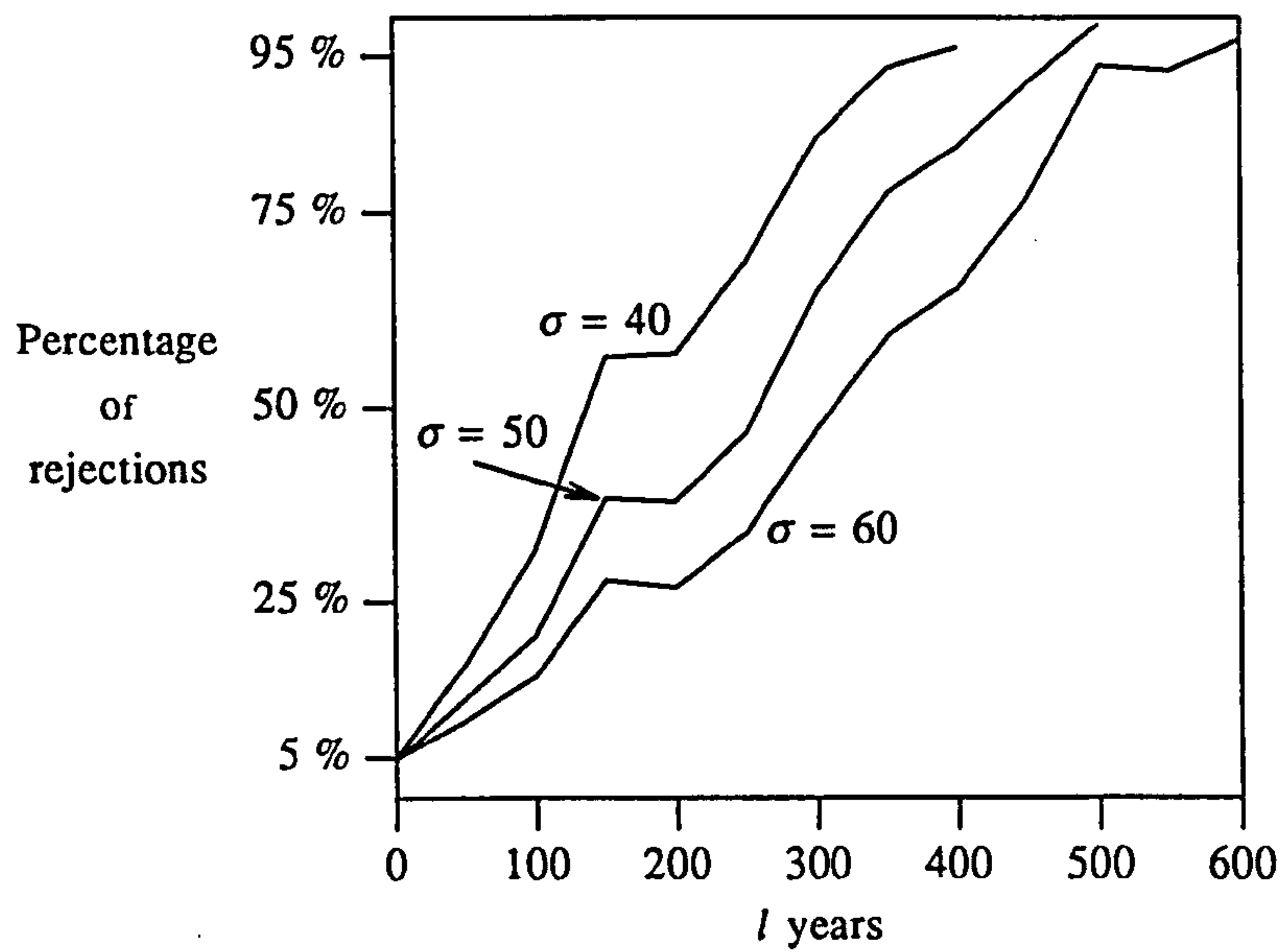


Figure 2.2
Percentage of rejections for the T -test,
for $\alpha = 0.95$, $m = 5$, $n = 5000$, $\beta = 4050$ and $\sigma = 40, 50$ and 60 .

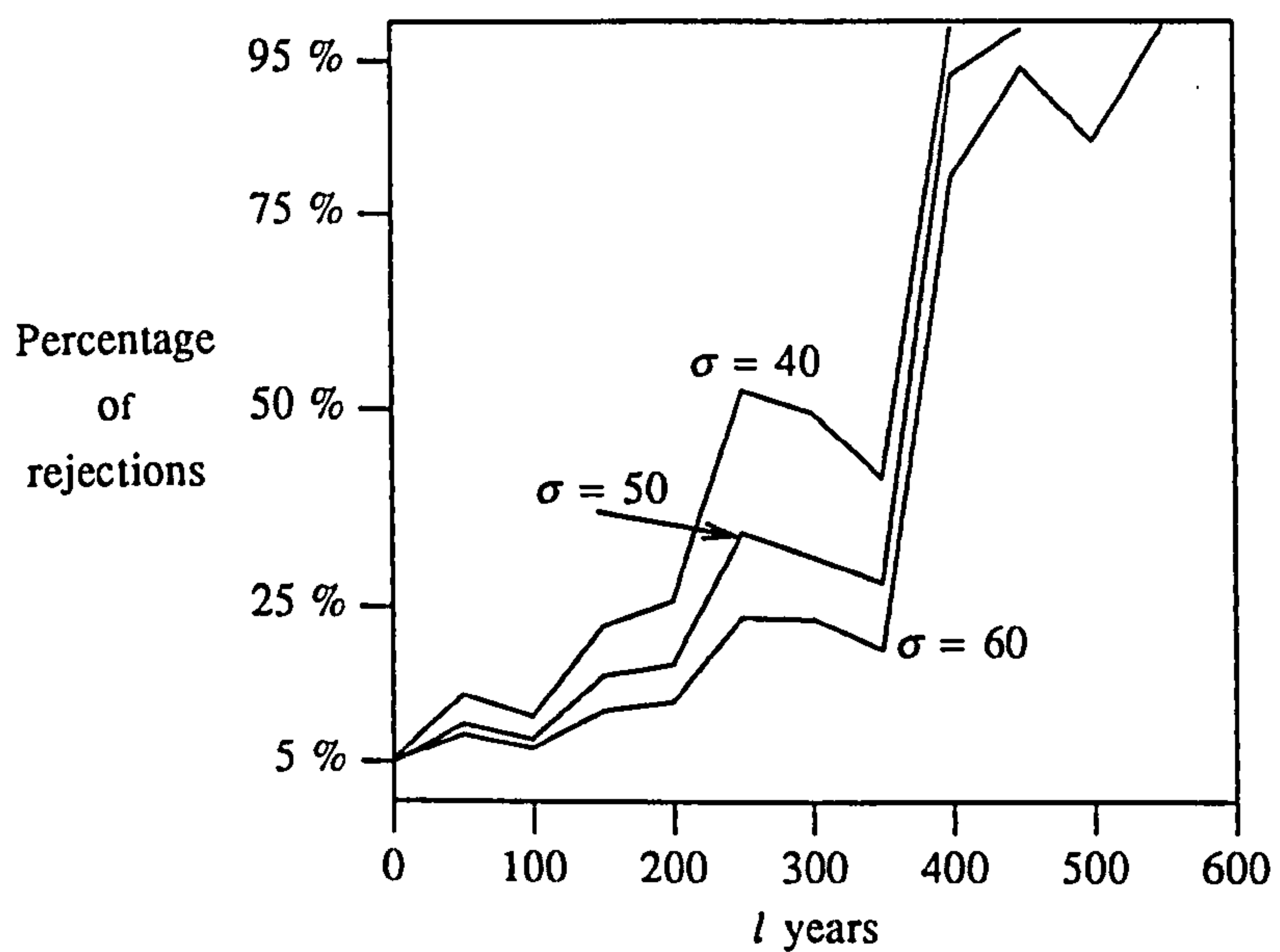


Figure 2.3
Percentage of rejections for the T -test,
for $\alpha = 0.95$, $m = 5$, $n = 5000$, $\beta = 4550$ and $\sigma = 40, 50$ and 60 .

"A probability test is given (...) to help distinguish whether, for example, an occupation floor was accumulated in an "instant" of time (several years, or sequential seasons, or less), or if the spread of radiocarbon dates probably indicates a real time spread such as many decades or centuries."

Suppose now that somehow we can overcome the problems of calibration mentioned before and that we propose a significance test to 'help distinguish whether (...) an occupation floor was accumulated in an "instant" of time'. The problem we foresee with such tests is that they are based, essentially, on a confidence level. Therefore, what we call 'instant' of time would depend on the choice of this confidence level. We are never given (in Ward and Wilson 1978 nor elsewhere) the relationship between such confidence level and the 'instant' of time (like its length, for example). This means that, what we call an 'instant' of time, will depend on what confidence level we are using, and we have no means to establish such dependency. This raises the following questions.

Should what we call a 'instant' of time be based on archaeological considerations?

Should we, at least, know how long our 'instant' of time is?

We believe that the answer to both of the above questions is 'yes' and, therefore, we find significance tests unsatisfactory within this context, at least in the way they have been presented up until now. Moreover, significance tests for a point null hypothesis (like H_0) have been criticised previously by several authors (Lindley 1957, Hays and Winkler 1970 chapter 7, Barnett 1973 chapter 5). In summary, to quote Berger (1985, p. 135)

"In the face of this overwhelming evidence that classical testing of a point null [hypothesis] is misleading, we must seek a better approach. Of course, we basically recommend the subjective Bayesian approach (...)"

2.4.4 Other major contributors

In a series of papers, Ottaway and her colleagues present a major contribution to the interpretation of radiocarbon determinations, by recognising the archaeological desire to summarise sets of determinations. The problem of summarising a set of radiocarbon determinations is quite common in archaeology. In this case the determinations are supposed to give evidence about the time-span of a given archaeological phenomenon. Statistical techniques are then needed to estimate this 'time-span' given the determinations.

Ottaway (1973) proposes a technique to summarise a given set of radiocarbon determinations $y_1 \pm \sigma_1, y_2 \pm \sigma_2, \dots, y_m \pm \sigma_m$. She avoids using averages (or the pooled mean) of determinations because 'one can see that this approach is meaningless'. The technique she proposes is called 'interquartile range' or 'dispersion diagrams', and consists of ordering the radiocarbon ages, y_j 's, in the radiocarbon scale and highlighting the first 25%, the middle 50% and the last 25% of the data, with predefined grey-shadowed boxes. This is called the 'interquartile range' for the determinations. Only the radiocarbon ages y_j 's are considered, and the standard deviations σ_j 's are ignored. This constitutes a purely graphical technique with no probabilistic model stated that has the following characteristics.

- (i) It does not consider the standard deviations reported for the radiocarbon determinations.
- (ii) It does not consider the archaeological context to which the radiocarbon determinations are related.

Later she proposes to 'correct' (calibrate) the radiocarbon determinations by finding the intercepts of the y_j 's with the calibration curve, and using these

intercepts (calendar years) to form a dispersion diagram as before, but now on the calendar scale.

There are several problems with this approach. A major concern is related to the arbitrary choice of the 25%, 50% and 75% intervals. Ottaway explains that (Ottaway 1973)

"it may be that the expansion to include two-thirds or three-quarters of the known dates will give empirically the more meaningful results, but this is a problem for the future. At the present the simplicity of the inter-quartile range has everything to commend it."

Another major concern with dispersion diagrams is that the standard deviations are not included in the analysis. This has the consequence that a radiocarbon determination has the same influence whether is low-precision ($\sigma = 60-80$) or high-precision ($\sigma = 15-20$). As a result of this, a set of high-precision determinations would have exactly the same dispersion diagram as a set of low-precision determinations if the radiocarbon ages reported (the y_j 's) happen to be the same. Ottaway defends her technique by saying (Ottaway 1973)

"the way the data are plotted in dispersion diagrams already contains within it a good reflection of the uncertainties of the data, and one does not need the apparent extra information given by the standard deviation (...)"

We do not agree with the above statement that the standard deviations reported represent 'apparent extra information'. If we neglect the σ_j 's we will be losing important, and not 'apparent', information. The σ_j 's are carefully assessed by the radiocarbon laboratories based on the imprecisions arising from the dating process (counting periods, pretreatments etc.) that may vary from sample to sample. Indeed, determination 1400 ± 70 is less precise than determination 1400 ± 40 given the standard deviation reported by the laboratory and this should always be

taken into account in the succeeding analysis and interpretation.

The next paper in the series is Ottaway (1986) which considers the 'calibration of groups of [radiocarbon] measurements', using data from a site in lower Bavaria, Germany. She proposes the use of dispersion diagrams but then poses the question,

"Is it more sensible to construct an interquartile range with uncalibrated dates (...) and then calibrate the range, or to calibrate each date and then construct an interquartile range? (...) the two methods give notably different values."

This question is difficult to answer, not having any clear statistical framework to refer to. Again in Ottaway (1986), the standard deviations are not considered in the analysis.

The last paper we review in this Section is Aitchison *et al.* (1991). This paper attempts to overcome the problems of the 'dispersion diagrams' by using the high-precision calibration curve and including the standard deviations reported for the radiocarbon determinations. The aim is to estimate 'the duration of an archaeological phenomenon' using a set of radiocarbon determinations. The approach used is based on the following principle

"Any formal approach must start with the assumption that there exists a frequency distribution (with respect to the historical time-scale) of *all* possible artefacts or materials from the phenomenon which might be sampled."

Here they try to define a 'population' or sample space from which the samples, and eventually, the radiocarbon determinations arise. We believe, though, that the definition of such a sample space is difficult if not impossible and introduces a number of questions. Does a broken artefact count as one individual in the population or as several (the materials resulting from the debris)? Do Aitchison

et al. refer to the materials that in the past existed or the ones that we can actually find in the present? Radiocarbon dating is only capable of dating some type of materials, does this represent a problem? What is the relationship of ‘*all* possible artefacts or materials’ to the duration of the ‘archaeological phenomenon’ to which the artefacts relate? Is it the same for *all* archaeological phenomena?

In order to implement their approach, Aitchison *et al.* (1991) propose to estimate the frequency distribution of ‘*all* possible artefacts or materials’ and then to estimate its lower and upper quartiles (similar to the dispersion diagrams mentioned earlier). They use a ‘kernel density estimation’ technique based on Silverman (1986). Given a set of radiocarbon determinations $y_1 \pm \sigma_1, y_2 \pm \sigma_2, \dots, y_m \pm \sigma_m$ they proceed as follows.

- (i) Solve the equations

$$y_j = \mu(\theta). \quad (2.5)$$

This results in a series of $\theta_{i,j}$ ’s where $y_j = \mu(\theta_{i,j})$ for all i (with proper ranges for all indices), see Figure 2.4. Recall that Equation 2.5 may have several solutions due to the non-monotonicity of the piece-wise linear calibration curve $\mu(\theta)$.

- (ii) Consider all the $\theta_{i,j}$ ’s found in (i) as a sample for kernel density estimation. The estimated density therefore belongs to the calendar scale and is taken to represent the distribution of ‘*all* possible artefacts or materials’.

One basic assumption of kernel density estimation is that the sample values are independent. Clearly the $\theta_{i,j}$ ’s are *not* necessarily independent. Referring to Figure 2.4, it is clear that given $\theta_{1,1}$ we know y_1 since

$$\mu(\theta_{1,1}) = y_1$$

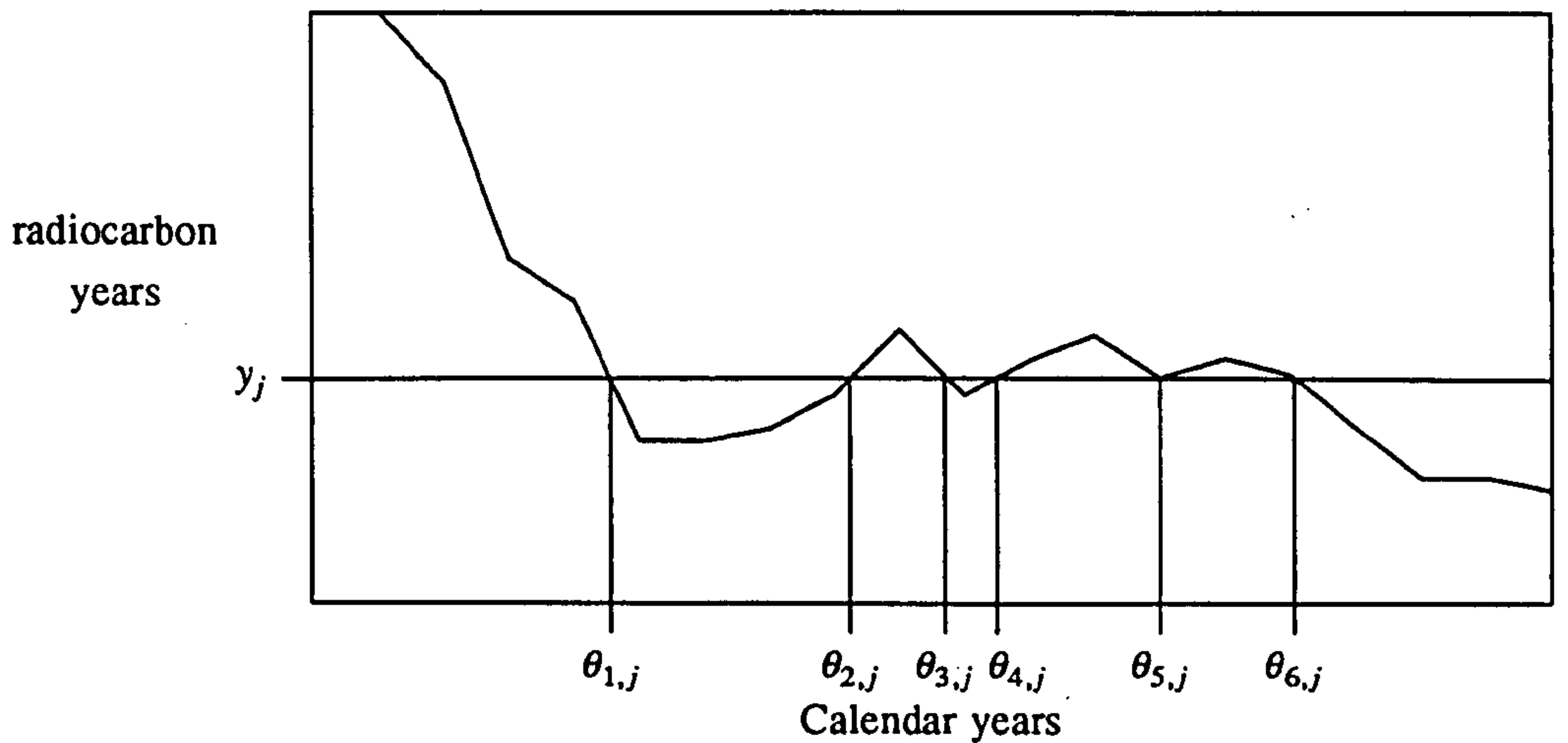


Figure 2.4
Solutions for the equation $y_j = \mu(\theta)$.

and we can then *calculate* the rest of the $\theta_{i,1}$'s.

We believe that Aitchison *et al.* (1991) is a work that is controversial both because of its archaeological assumptions and because the $\theta_{i,j}$'s do not necessarily satisfy the statistical assumptions required. In addition, no provision is made for further archaeological considerations to be incorporated into the statistical analysis, like archaeological phases, floating chronologies or known relationships between the dates. Estimating the duration of a particular archaeological phenomenon using radiocarbon is a very important problem in the interpretation of radiocarbon determinations. In Section 4.3.5 we present a novel approach to this problem using our Bayesian method.

2.4.5 Archaeological wiggle matching

The technique of archaeological wiggle matching (AWM) was presented in Weninger (1986) and Pearson (1986) and has been used in some publications including Baillie and Pilcher (1988), Manning and Weninger (1992), Baillie

(1990) and Clymo *et al.* (1990). Other earlier attempts to tackle problems of AWM without using the high-precision calibration data can be found in Clark and Renfrew (1972), Clark and Sowray (1973) and Clark and Morgan (1983). The latter group of papers are based on a technique to date floating tree-ring chronologies that was formally presented by Clark (1979) and reviewed in Section 2.3.4. Here we concentrate on the former group of papers which make use of the high-precision calibration data.

In the explanation of the problems considered by AWM, Weninger (1986) states that

"Cultural phenomena known by Comparative Stratigraphy to be sequential appear to overlap on the radiocarbon scale. A rectification is not achieved by calibrating either the dates or the radiocarbon scale. Calibration does not put dates into their correct order and it does not guarantee that the real sample age is known."

The aim, therefore, is to combine known archaeological (relative) chronologies with radiocarbon determinations to obtain absolute chronologies. In other words, to fix in time the relative chronologies using radiocarbon. Their technique consists of the following steps.

- (i) Ordering the radiocarbon determinations according to the archaeological chronology ie. the object dated by determination $y_j \pm \sigma_j$ is known to be earlier than the object dated by determination $y_{j+1} \pm \sigma_{j+1}$ (but not taking into account the the values of $y_j \pm \sigma_j$ and $y_{j+1} \pm \sigma_{j+1}$, only the archaeological information).
- (ii) Plotting $y_j - \sigma_j$, y_j , $y_j + \sigma_j$ along evenly spaced vertical lines, the gap between the lines being n calendar years.

- (iii) Using a graph of the high-precision calibration curve, with the same scale for radiocarbon years (y axis) and calendar years (x axis) as for the plot in (ii) and attempt to find a section on the curve that would match (at least approximately) the pattern of the radiocarbon determinations plot from (ii). This could be done by using a transparent slide for the plot in (ii).
- (iv) Repeating steps (ii) and (iii) with different gaps of n calendar years until a 'satisfactory' match is found.

This 'satisfactory' match then provides an estimate for the calendar age for each of the dated objects. Clearly these estimates include the *a priori* relative chronological information available.

Weninger's opinion about his own AWM method is that

"AWM is the art of making good guesses. I would define "good guesses" as being hypotheses reproducible in an inter-archaeologists study group."

There are obvious difficulties in comparing a highly subjective technique with other more formal techniques, methods or frameworks. The principal problem is comparing how good our 'good guesses' are. That is, what measure should we use to compare one wiggle match with another? We do sympathise with Weninger's concern about the necessity of introducing archaeological considerations into the analysis of radiocarbon data. However, this *can* be done in a consistent way resulting in reproducible techniques by an inter-archaeologists study group, with the help of a trained (Bayesian) statistician. We therefore paraphrase Weninger by saying that 'Bayesian statistics is, partly, a methodology to perform consistent and coherent good guesses'. Finally, we agree with Weninger's statement that

"Further research should turn up with a better mathematical approach."

Pearson (1986) presents a more mathematical approach to the problem of AWM. He concentrates on dating floating tree-ring chronologies. In this case the radiocarbon determinations arise from samples of tree-rings where the number of rings between successive samples is known. Therefore, if $\theta_1, \theta_2, \dots, \theta_m$ are the associated calendar years for the determinations we have

$$\theta_j = \theta_{j-1} + l_j$$

where l_j is known. Here we suppose that $\theta_m > \theta_{m-1} > \dots > \theta_2 > \theta_1$ and thus θ_1 represents the calendar age for the outermost (youngest) ring of the tree-ring chronology. Since the l_j 's are known it is only needed to estimate θ_1 and the rest of the θ_j 's can then be calculated since $\theta_j = \theta_1 + \sum_{k=2}^j l_k$.

Pearson (1986) uses a technique based on least squares to estimate θ_1 . The basic idea is to minimise the sum of squares $\sum_{j=1}^m \left\{ y_j - \mu \left(\theta_1 + \sum_{k=2}^j l_k \right) \right\}^2$ as a function of θ_1 . Pearson assumed that $\sigma_j^2 = \sigma^2$ for all determinations. However, if the σ_j^2 's were different, the least squares method could easily be extended to estimate θ_1 by finding the minimum of

$$S(\theta_1) = \sum_{j=1}^m \frac{1}{\sigma_j^2} \left\{ y_j - \mu \left(\theta_1 + \sum_{k=2}^j l_k \right) \right\}^2.$$

As explained in Pearson (1986), one can see that, given θ_1 , $S(\theta_1)$ has a chi-square distribution (using the common calibration model for radiocarbon determinations presented in Chapter 1) and therefore confidence intervals can be calculated. Unfortunately, this technique can only be used when the l_k 's are known and seems difficult to generalise to more complicated cases when little or no information is available about the l_j 's. Therefore there is a need to seek a more general approach to tackle the problems of AWM. This will be addressed in

Section 5.3 where we develop an approach to the problem using our (Bayesian) method.

2.4.6 'El Castillo' determinations

Harrison and Wainwright (1991) analyse a set of radiocarbon determinations arising from 'El Castillo', a site in Teruel, Spain. They explain that

"A big problem with using radiocarbon determinations is the difficulty encountered with converting the irregular recalibrated values into calendar dates and of combining them. This paper takes six new determinations and by including the archaeological evidence, aims to produce simple estimates."

The problem is to incorporate a chronological phases structure in the analysis, using a set of six radiocarbon determinations. That is, there is a series of archaeological phases identified in the site which are known to have some relative chronological ordering. For each phase there are one or more determinations associated with it. On the basis of some archaeological considerations Harrison and Wainwright use only four of these determinations. One further determination is rejected because it 'falls out of chronological sequence' and thus the analysis is performed with only three determinations which we denote by $y_1 \pm \sigma_1, y_2 \pm \sigma_2, y_3 \pm \sigma_3$.

The calendar dates associated with these determinations, $\theta_1, \theta_2, \theta_3$, are then taken to predate and postdate the boundaries for three consecutive phases in the site. That is, θ_1 predates the beginning of phase I and θ_2 postdates its ending, θ_2 predates the beginning and θ_3 postdates the ending of phase II and θ_3 predates the beginning of phase III. Thus, using years BP, it is known that

$$\theta_1 > \theta_2 > \theta_3.$$

Harrison and Wainwright's (1991) first approach to dating the three archaeological phases in 'El Castillo' can be explained in the following way. What is required is the most likely period in which the phases occurred. Since phase I occurred between θ_1 and θ_2 then, given the length, v , of this phase, the probability that the phase occurred in any period of time α to $\alpha - v$ is equal to

$$P_1(\alpha, v) = P[\theta_1 > \alpha, \alpha - v > \theta_2 \mid y_1, y_2].$$

Harrison and Wainwright assume that this probability is equal to $P[\theta_1 > \alpha \mid y_1]P[\alpha - v > \theta_2 \mid y_2]$. Using basically the same calibration procedure explained in Chapter 1, they obtain $P[\theta_j < \alpha \mid y_j]$ for $j = 1, 2, 3$, and hence $P_1(\alpha, v)$ can be calculated. Therefore Harrison and Wainwright assume, albeit implicitly, that given the determinations y_1 and y_2 , θ_1 and θ_2 are independent. It is unlikely that this is the case since, *a priori*, we know that $\theta_1 > \theta_2$ (they do not discuss this problem). Of even greater concern is their method for estimating the most likely period in which phase I occurred. They find α^* and v^* such that

$$P_1(\alpha^*, v^*) = \max_{\alpha, v} P_1(\alpha, v).$$

Although this approach might be intuitively appealing it presents serious problems. Since $P[\alpha - v > \theta_2 \mid y_2] < P[\alpha > \theta_2 \mid y_2]$ then it is straightforward to see that $v^* = 0$, and therefore the most likely period of time for phase I would be of no length. Surprisingly, Harrison and Wainwright give 2057 BC and 10 years as estimates for α^* and v^* (a similar analysis is given for phase II and no analysis for phase III).

These authors recognise that the estimates for the time in which the three phases occurred should include the three determinations simultaneously, since it is known that $\theta_1 > \theta_2 > \theta_3$. They therefore introduce a second approach to the problem saying that

"To rectify this, one can limit the time ranges used, and combine all dates to give a delimited three-way probability"

which is defined as

$$P_3(\Delta t) = \max_{\alpha, v_1, v_2} (P[\alpha > \theta_1 > \alpha - \Delta t, \alpha - v_1 - \Delta t > \theta_2 > \alpha - v_1 - 2\Delta t, \\ \alpha - v_1 - v_2 - 2\Delta t > \theta_3 > \alpha - v_1 - v_2 - 3\Delta t \mid y_1, y_2, y_3]).$$

Again, it is assumed that given the determinations, θ_1 , θ_2 and θ_3 are independent and $P[\theta_j < \alpha \mid y_j]$ is used to calculate $P_3(\Delta t)$. This represents the probability that θ_1 , θ_2 and θ_3 belong to three consecutive periods of time, each of length Δt , and separated by v_1 and v_2 . Therefore, according to the first analysis, phase I should have occurred between $\alpha - \Delta t$ and $\alpha - v_1 - \Delta t$ and phase II between $\alpha - v_1 - 2\Delta t$ and $\alpha - v_1 - v_2 - 2\Delta t$. According to the second analysis, phase I is claimed to be between α and $\alpha - \Delta t$, phase II between $\alpha - v_1 - \Delta t$ and $\alpha - v_1 - 2\Delta t$ and phase III between $\alpha - v_1 - v_2 - 2\Delta t$ and $\alpha - v_1 - v_2 - 3\Delta t$. Given a 'time-span' Δt one can find α^* , v_1^* and v_2^* to maximise the above probability to obtain $P_3(\Delta t)$. Using *ad hoc* graphical methods Δt is estimated to be 40 years and hence the three periods of time for the phases are obtained using the corresponding values for α^* , v_1^* and v_2^* that give $P_3(40)$.

Although the techniques proposed by Harrison and Wainwright (1991) are controversial and perhaps contradictory, we believe that their basic principle is highly relevant. As stated above, the intention is to combine the known archaeological information available (chronological ordering in the phases, phase lengths) with the radiocarbon determinations. This idea is emphasised throughout the paper recognising the fact that, otherwise, a sensible interpretation of the 'El Castillo' determinations cannot be satisfactorily achieved. Using the Bayesian framework, combining *a priori* archaeological information and radiocarbon determinations can be done in an explicit and consistent manner. In fact, using our

Bayesian methodology, in Section 5.3 we study and propose solutions to problems similar to the one discussed here.

Moreover, Harrison and Wainwright reject one determination because its calibrated distribution falls outside of the time periods indicated by other distributions associated with the same context ('falls out of chronological sequence'). However, this rejection is based only on intuitive ideas and *ad hoc* informal comparisons that cannot be consistently reproduced in other analyses. Again, using the Bayesian framework, the accordance (or discordance) of a group of determinations can be measured and the rejection of a determination can be formally included as part of the analysis. Using the extended statistical framework to be developed in Chapter 6, a full analysis of this type can be performed.

2.4.7 Bayesian approaches

In this final Section we review publications concerning the interpretation of radiocarbon determinations that utilise the Bayesian framework. These publications are dominated by an interest in new modelling approaches, with the aim of introducing specific considerations into the statistical analysis depending on the archaeological problem at hand. Based upon complex models and sound statistical techniques, complicated archaeological dating problems are tackled. The Bayesian framework is then used to combine sets of radiocarbon determinations with relevant archaeological (chronological) information in a consistent manner. The first paper we study is Helskog and Schweder (1989), they explain that

"The problem at hand is to estimate the number of contemporaneous houses (units) at any given time t . [from a site in Norway]"

They have m radiocarbon determinations arising from charcoal found in m houses in the site. The radiocarbon determinations were calibrated using the Ralph *et al.* (1973) procedure explained in Section 2.3.3. Helskog and Schweder follow the recommendations of Ralph *et al.* and consider the *calibrated* determinations to be normally distributed (see below). In the paper, only the calibrated results are given, the uncalibrated determinations are not given. (We have discussed the inadequacies of such an approach in Section 2.3.3.)

The procedure that these authors use is basically as follows. Let the occupation period of each house be equal to L years. Let τ_j be the date when house j was abandoned, u_j the deposition time of the j^{th} charcoal sample taken from house j and $y_j \pm \sigma_j$ the determination arising from that sample. Then Helskog and Schweder (1989) state that

$$y_j \mid u_j, \tau_j \sim N(u_j, \sigma_j^2). \quad (2.6)$$

Since there is no further information about the time of deposition for each sample within the occupation period of the houses, u_j is assumed uniform with

$$u_j \mid \tau_j \sim U(\tau_j, \tau_j + L)$$

and the prior distribution of τ_j is uniform within some wide margins. (Note that, since we are using years BP, τ_j , the date of abandonment of house j , is *smaller* than the beginning of the occupation $\tau_j + L$.) Applying Bayes' theorem, Helskog and Schweder obtain

$$f(\tau_j \mid y_j).$$

That is, the posterior distribution of the date of the abandonment of house j given the radiocarbon determination performed for that house. Using these distributions they calculate the expected number of houses occupied at time t (for fixed L), $N(t \mid y)$. They do this by noticing that, given $\tau = (\tau_1, \tau_2, \dots, \tau_m)$, the number of

houses occupied at any time t is

$$N(t, \tau) = \sum_{j=1}^m I(t)_{(\tau_j, \tau_j+L)},$$

where $I(t)$ is the indicator function. That is, a house j is occupied at a time t if $t \in (\tau_j, \tau_j+L)$. Therefore, the expected number of houses occupied at time t is

$$N(t | y) = \int N(t, \tau) f(\tau | y) d\tau.$$

Therefore, since $f(\tau | y) = \prod_{j=1}^m f(\tau_j | y_j)$ and $I(t)_{(\tau_j, \tau_j+L)}$ (viewed as a function of τ_j) is equal to $I(\tau_j)_{(t-L, t)}$, we have

$$N(t | y) = \sum_{j=1}^m \int_{t-L}^t f(\tau_j | y_j) d\tau_j.$$

Helskog and Schweder then calculate and plot $N(t | y)$ for $L = 50, 75$ and 100 years. They conclude that the maximum number of houses occupied at the same time is between 1.8 and 3.5 , although this is obviously extremely dependent on the value of L . We believe that using a fixed value of L is a severe constraint which could be avoided by introducing a prior distribution for L and then integrating out L to calculate $N(t | y)$.

As we stated above, we are only given 'calibrated' determinations. If we were given the original uncalibrated determinations, the model could be extended by changing Equation 2.6 to

$$y_j | u_j, \tau_j \sim N(\mu(u_j), \sigma_j^2),$$

where y_j is now the original (uncalibrated) radiocarbon determination, measured in radiocarbon years (and μ is the piece-wise linear calibration curve).

Helskog and Schweder's (1989) paper is important because it clearly introduces archaeological considerations into the statistical analysis. We believe that

only in this way can the problem of estimating the number of contemporaneous houses be achieved. In this dating problem it is crucial to introduce into the analysis considerations about,

(i) the longevity of the houses

and

(ii) the relationship between deposition time of an object and its radiocarbon determination.

This can only be done by proper statistical modelling and the introduction of some type of prior information. Only the Bayesian framework provides a means of addressing the archaeological questions. (We continue to study this problem in Section 5.4.)

The next paper we review is Naylor and Smith (1988). This paper represents a major contribution to the Bayesian interpretation of radiocarbon determinations. The archaeological problem studied is from the Iron Age hillfort in Danebury, England. There are a total of 65 radiocarbon determinations and 'each of the 65 samples is associated with a pottery shard or fragment'. The problem was considered previously by Orton (1983) who was unable to include the calibration process in the analysis. Naylor and Smith (1988) explain that,

"The 65 radiocarbon dates are thus accepted by the archaeologists as the dates for the 65 associated pottery fragments. (...) individual fragments are regarded as being classified into one of four phases, which we refer to as Ceramic Phases 1-4. These phases are regarded as abutting nonoverlapping periods of stylistically consistent production (...)"

They use $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_5)$ to represent the beginning and end for each of the four phases with the assumption that

$$\alpha_1 > \alpha_2 > \alpha_3 > \alpha_4 > \alpha_5,$$

in other words, the phases do not overlap. Using $\theta_{i,j}$ to 'denote the actual date of manufacture of the original pot' j assigned to ceramic phase i , they assume $f(\theta_{i,j} | i, \alpha) = f(\theta_{i,j} | \alpha_i, \alpha_{i+1})$. They explain that this distribution 'could reflect assumptions that seem appropriate about the relative rate of production of pottery (...)'. They choose $f(\theta_{i,j} | \alpha_i, \alpha_{i+1})$ to be a uniform distribution on (α_i, α_{i+1}) , that is

$$f(\theta_{i,j} | \alpha_i, \alpha_{i+1}) \propto \begin{cases} \frac{1}{\alpha_{i+1} - \alpha_i} & \text{if } \alpha_{i+1} > \theta_{i,j} > \alpha_i \\ 0 & \text{otherwise.} \end{cases}$$

Finally they assume that

$$y_{i,j} \sim N(\mu(\theta_{i,j}), \sigma_{i,j}^2). \quad (2.7)$$

To calculate the posterior distribution of α they use a 'method based on interactive rescaling of Gauss-Hermite Cartesian product rules', using their own computer software. The posterior distribution of α is then used to give evidence about when the beginning and end of the different ceramic phases occurred. Therefore in their analysis Naylor and Smith (1988) successfully combined archaeological information (ceramic stylistic phases) with the radiocarbon determinations available, to estimate the ages of occurrence of the different ceramic phases identified in Danebury.

Naylor and Smith's (1988) paper contains two technical errors. The first one is that they consider the year 0 BP as 1983 AD. This is erroneous and should be corrected to 1950 AD. The second error is that they use an old calibration data set instead of the high-precision calibration data available since 1986.

Besides these two technical errors there are a number of further questions that should be considered. It is unclear what the relationship is between the pottery and the samples dated ('charcoal, grain and animal bone'). The model states that each radiocarbon determination is associated with a ceramic pot, this in turn is associated with a ceramic phase. What is the nature of this relationship? How are we going to infer the relationship between the radiocarbon determinations and the ceramic phases? In the paper there is no clear discussion about this problem but it is implicitly included in the modelling assumption in Equation 2.7, by stating that the determinations $y_{i,j}$'s have an associated calendar year $\theta_{i,j}$ equal to the 'manufacturing of the original pot'. This implies that in the year the organic matter in object j died then, the pot it relates to was manufactured. This could well be a reasonable assumption but was not clearly explained nor discussed in the paper.

To finish with this bibliographical review we are going to study three papers. Litton and Leese (1991), Buck *et al.* (1991) and Buck *et al.* (1992). These three papers are closely related to each other, as well as to Naylor and Smith (1988).

Litton and Leese (1991) present the basic modelling ideas of Naylor and Smith (1988) in a clearer archaeological exposition. They consider the basic model

$$y \sim N(\mu(\theta), \sigma^2),$$

where θ is the associated calendar age for determination $y \pm \sigma$ and $\mu(\cdot)$ is the piece-wise linear calibration curve. The calibration of the radiocarbon determinations is therefore explicitly included in the model. Buck *et al.* (1991) and Buck *et al.* (1992) also use this model.

The combination of Naylor and Smith (1988) and Litton and Leese (1991) leads to a new approach to the statistical analysis and interpretation of sets of

radiocarbon determinations that is based on the Bayesian framework. This new approach is followed up by Buck *et al.* (1991) and Buck *et al.* (1992). The former is a paper directed at archaeologists and includes a basic exposition of the Bayesian framework with some illustrative examples. They present a very simple example from a site called 'Skara Brae' to illustrate the Bayesian methodology. In this case they have four events represented by the calendar years θ_1 , θ_2 , θ_3 and θ_4 which, *a priori*, have the ordering

$$\theta_1 > \theta_2 \geq \theta_3 > \theta_4.$$

They introduce this information in the analysis through the joint prior distribution $f(\theta_1, \theta_2, \theta_3, \theta_4)$. Then they calculate the posterior distributions of each of the events given the radiocarbon determinations. Therefore these posterior distributions not only consider the radiocarbon determinations but also the above prior information about the chronological ordering of θ_1 , θ_2 , θ_3 and θ_4 .

Buck *et al.* (1992) is a slightly more technical paper directed mainly at statisticians and archaeological scientists. They introduce the use of the 'Gibbs sampler' technique to approximate the posterior distributions of the parameters. In this paper a further example from a site called 'Runnymede Bridge' is presented. This example has four events represented by $\theta_1, \theta_2, \theta_3, \theta_4$ but, *a priori* it is known that $\theta_1 > \theta_2$, $\theta_1 > \theta_3$ and $\theta_1 > \theta_4$. This information is then explicitly introduced into the analysis via the specification of the joint prior distribution $f(\theta_1, \theta_2, \theta_3, \theta_4)$. The posterior marginal distributions $f(\theta_j | y)$ are calculated using the 'Gibbs sampler' technique (see Section 4.3.4). Again, the resulting distributions will not only be based on the radiocarbon determinations but also on the archaeological information available (the *a priori* chronological ordering of the θ_j 's).

As a further, though more complicated example, Buck *et al.* (1992) reconsider the Danebury radiocarbon determinations with a larger archaeological explanation. They study the model involved and relaxed the assumption of abutting phases, as used by Naylor and Smith (1988). Possible inadequacies of the abutting phases assumption clearly emerge by comparison with the case of overlapping phases. In the latter model each ceramic phase i is represented by its beginning α_i and its end β_i and they assume that

$$\alpha_i > \beta_i$$

(in the abutting phases case it is assumed that $\beta_i = \alpha_{i+1}$). Then they use the Gibbs sampler to obtain the posterior distributions for the α_i 's and β_i 's. We believe that this represents a more satisfactory approach to the 'Danebury' problem that could give archaeologists a clearer insight.

Taken together, Naylor and Smith (1988), Litton and Leese (1991), Buck *et al.* (1991) and Buck *et al.* (1992) represent a novel approach to the interpretation of sets of radiocarbon determinations. The basis of this (Bayesian) approach consists of the following points.

- (i) Understanding the archaeological dating problem at hand.
- (ii) Establishing a basic statistical model that includes calibration.
- (iii) Building a mathematical (statistical) model to relate the archaeological dating problem with the radiocarbon determinations and including in the analysis relevant *a priori* information available.
- (iv) Combining (ii) and (iii) consistently using the Bayesian framework to obtain results interpretable within the archaeological dating problem in (i).

In this thesis we will adopt this same basic approach, improving it with our own original ideas and techniques, as we believe that only in this way can sound interpretations of radiocarbon determinations be achieved. In all the examples reviewed in this Section it has been crucial to include in the analysis the relevant available archaeological (chronological) information and combine it with the radiocarbon determinations. The Bayesian framework for statistical inference provides a consistent method for doing precisely that.

2.4.8 Discussion

From the bibliographical review concerning publications relevant to the interpretation of radiocarbon determinations, one immediate conclusion can be drawn: there is no unified procedure for analysing and interpreting a set of radiocarbon determinations. Since the advent of radiocarbon dating, various methods have been proposed, but none has been universally accepted by the radiocarbon community. Moreover, the methods or techniques used tend to be rather *ad hoc* and directed at very specific examples or, at best, have only limited scope. From the papers reviewed above we can identify the following typical problems in the interpretation of radiocarbon determinations.

- (i) When and how to average a set of determinations to give a single date. This represents a very common problem that has been studied by several authors of which we have mentioned Spaulding (1958), Polach and Golson (1968), Leach (1972), Law (1975), Long and Rippeteau (1974) and Ward and Wilson (1978).
- (ii) How to summarise a set of radiocarbon determinations to give evidence about the time-span or duration of a given archaeological phenomenon.

Some authors have studied this problem and of these we mentioned Ottaway (1973), Ottaway (1986) and Aitchison *et al.* (1991).

- (iii) How to obtain an absolute chronology of events from known relative chronological orderings (or 'fix' a floating chronology) using radiocarbon dating. This is a very common and important problem that has been studied by several authors. Of these we have mentioned Clark and Renfrew (1972), Clark and Sowray (1973), Clark (1979), Clark and Morgan (1983), Weninger (1986), Pearson (1986), Baillie and Pilcher (1988), Manning and Weninger (1992), Baillie (1990), Clymo *et al.* (1990) and Buck *et al.* (1992).
- (iv) How to date archaeological phases known to have some chronological order using radiocarbon dating. This problem has been addressed by several authors of which we mentioned Harrison and Wainwright (1991), Naylor and Smith (1988) and Buck *et al.* (1991).

We have reviewed and criticised the different approaches used by all of these authors and of these, the Bayesian approach seems to be the most promising. However, problems (i) to (iv) can all be classified as being within the statistical analysis and interpretation of radiocarbon determinations and as yet no global methodology has been developed to tackle them. This motivates the major objective of this thesis: to, use the Bayesian framework and develop a general statistical framework for the interpretation of radiocarbon determinations. Such a framework is given in Chapter 4 and then used in the remainder of the thesis to provide solutions to problems of the type (i) to (iv). (More specifically, problem (i) is principally studied in Section 4.4, problems (ii) and (iii) in Sections 5.2 and 5.3 respectively and in Section 5.6 we address problem (iv).) We show then that all of these problems (and several others) can be studied within a global framework

for the analysis and interpretation of radiocarbon determinations.

2.5 Discussion

Given the wide range of problems and the number of papers reviewed here, it is not difficult to realise that the area of radiocarbon dating and in particular the problem of interpreting radiocarbon determinations, is influenced by several, perhaps, contradictory schools of thought. However, we can conclude three major points that have generally been accepted.

- (i) The standard deviations now reported do reflect the uncertainties in the radiocarbon dating process (before calibration), although there is always the risk of a mistake in the radiocarbon dating process that could generate an erroneous determination (outlier).
- (ii) There is generally accepted calibration data, with (possibly) global applicability (the high-precision calibration data).
- (iii) There is no general methodology for the interpretation of radiocarbon determinations.

In relation to point (i), we have assumed throughout good reliability of the radiocarbon dating process, considering radiocarbon determinations to be good estimates for the radiocarbon age of the samples dated. However, we consider the outlier problem to be a crucial element in the analysis and interpretation of radiocarbon determinations and this is studied in Chapter 6 where a novel approach to the problem is explained.

With respect to point (ii), the calibration data we use is the high-precision calibration data. However, there is still some controversy about the process of calibration to be used. Several techniques have been proposed but none achieved a general acceptance. In Chapter 3 we analyse this problem, developing a novel approach for the calibration of radiocarbon determinations.

In relation to point (iii), the interpretation of radiocarbon determinations is an even more controversial field. Several techniques have been proposed but an agreed core looks very remote. However, we have stressed the fact that in order to achieve a proper interpretation of radiocarbon determinations we need to include in the statistical analysis other sorts of information apart from the determinations themselves. We need to include considerations about the type of dating problem we are working with and its relation with the determinations. So far, only the techniques outlined in Naylor and Smith (1988), Litton and Leese (1991), Buck *et al.* (1991) and Buck *et al.* (1992) allow this possibility. It is achieved by using a satisfactory calibration procedure and relying on a consistent framework for statistical inference (the Bayesian framework).

We believe that to aim at an ultimate solution to the statistical analysis and interpretation of sets of radiocarbon determinations would be an illusion. Instead, in this thesis we consider a more skeptical, or rather humble, possibility. What we are proposing is, based on the Bayesian method, to build a general framework to tackle the most common statistical problems in the interpretation of radiocarbon determinations. The framework represents a possibility - a tested possibility.

Chapter 3

Calibration

As discussed in Section 1.2, the model currently proposed for the calibration of a radiocarbon determination $y \pm \sigma$ associated with the calendar year θ (the year in which the organic matter in the the object to be dated died) is

$$y \mid \theta \sim N(\mu(\theta), \sigma^2),$$

where $\mu(\theta)$ is the piece-wise linear calibration curve. Given the high-precision calibration data (see Section 1.1.2)

$$z = ([t_1, (x_1 \pm \sigma_1)], [t_2, (x_1 \pm \sigma_2)], \dots, [t_n, (x_n \pm \sigma_n)]),$$

we have $\mu(\theta)$ defined to be

$$\mu(\theta) = x_k \left(\frac{\theta - t_{k-1}}{t_k - t_{k-1}} \right) + x_{k-1} \left(\frac{t_k - \theta}{t_k - t_{k-1}} \right); \quad t_k > \theta \geq t_{k-1}, \quad k = 1, 2, \dots, n,$$

where $t_0 = 0$ and $x_0 = 0$. An obvious limitation of the above model is that we are neglecting any uncertainty in the calibration data. In other words, we are assuming that the x_k 's are not subject to any sampling error and therefore that $\mu(\theta)$ is known exactly. The rationale used to justify this is that, since they are based on replicated samples, the variances (σ_k^2 's) are small in comparison with the variance of an individual radiocarbon determination reported from a laboratory (typical values for the σ_k 's are 10 to 20 years, whereas a routine determination has a standard deviation of 40 to 80 years). Nevertheless, as radiocarbon techniques improve we may be working with high-precision determinations with standard deviations of 15 years or less. In such circumstances it may be important to include the σ_k^2 's within the calibration process. This problem has not been

tackled before within the Bayesian framework and is the subject of this Chapter.

3.1 Modelling the calibration curve

As has been previously pointed out, it is generally accepted by the radiocarbon community that the radiocarbon calibration curve should follow the ‘jumps’ in the calibration data and thus have numerous ‘wiggles’. It is considered not to be linear nor monotonic nor smooth, but is believed to be continuous. Thus, something substantially more complicated than a piece-wise linear fit or a cubic-spline may be needed to obtain a reliable modelling approximation. The approach we are going to take is to model the calibration curve as a stochastic process. Firstly, we denote the calibration curve by $r(\theta)$ and leave $\mu(\theta)$ exclusively for the particular case of the piece-wise linear calibration curve, as defined in Equation 1.3. Secondly, using some basic assumptions we will model $r(\theta)$ and, using some experimentally derived data, we will develop an alternative calibration process.

From the radiocarbon calibration data we know that $x_k \pm \sigma_k$ is a radiocarbon determination, and is associated with the calendar year t_k . This means that (as in Section 1.2)

$$r(t_k) \mid x_k, \sigma_k^2 \sim N(x_k, \sigma_k^2). \quad (3.1)$$

If we ignore the variances, σ_k^2 , then we have $r(t_k) = x_k$ (as it is the case for $\mu(t_k)$). Now instead of knowing $r(t_k) = x_k$, we think of $r(t_k)$ as a random variable normally distributed with mean x_k and variance σ_k^2 . The problem is then to extend the definition of $r(\theta)$ for $\theta \neq t_k$ that is, for the calendar years between the knots. This will be done in the next Section.

3.1.1 The process $r(\theta)$

From Equation 3.1 we know the distribution of $r(t_k)$ given the calibration data. We then need to model $r(\theta)$ between the knots of the calibration data in order to extend its definition to all of the calendar scale. Now, the intervals between the knots on the calendar scale are about twenty years. To model the behaviour of $r(\theta)$ in such intervals we will utilise information regarding the annual changes in the atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio. We make the assumption that $r(\theta)$ depends only on the two nearest knots, that is

$$f(r(\theta) \mid z) = f(r(\theta) \mid z_k, z_{k-1}); t_k > \theta \geq t_{k-1}, k = 1, 2, \dots, n. \quad (3.2)$$

In other words, given the calibration data z , the calibration curve in the interval t_k to t_{k-1} depends only on (x_k, σ_k^2) and $(x_{k-1}, \sigma_{k-1}^2)$.

Let $M(\theta)$ be the atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio at year θ BP. If an object has an associated calendar age θ , then using the 'law of radioactive decay' (Equation 1.1) the $^{14}\text{C}/^{12}\text{C}$ ratio, M , in the object at the present time is given by

$$M = M(\theta) \exp\left(-\frac{\theta}{\tau}\right). \quad (3.3)$$

On the other hand, the radiocarbon age for an object with a current $^{14}\text{C}/^{12}\text{C}$ ratio M is

$$t = -\tau \ln\left(\frac{M}{M_0}\right), \quad (3.4)$$

where M_0 is a constant known as the modern standard.

The idea now is to use information we have about the process $M(\theta)$, the atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio through time, to find a model for $r(\theta)$. Given the calibration curve $r(\theta)$, we note that $t = r(\theta)$. Combining this with Equations 3.3 and 3.4 we obtain

$$r(\theta) = \theta - \tau \ln \left(\frac{M(\theta)}{M_0} \right),$$

and hence that

$$r(\theta-1) = r(\theta) - 1 - \tau \ln \left(\frac{M(\theta-1)}{M(\theta)} \right). \quad (3.5)$$

The interpretation of Equation 3.5 is that $r(\theta-1)$ is equal to $r(\theta)$ minus one year minus some 'distortion' represented by the term

$$\tau \ln \left(\frac{M(\theta-1)}{M(\theta)} \right),$$

which is dependent on the relative change of the atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio from year θ to year $\theta-1$. Within the radiocarbon literature, a common way to measure $M(\theta)$ is to use the value $\Delta^{14}\text{C}$, measured in 'parts per mil', which is defined (see Stuiver and Polach 1977) as

$$\Delta^{14}\text{C} = \left(\frac{M(\theta)}{M_0} - 1 \right) 10^3.$$

If we let

$$\Delta(\theta) = \left(\frac{M(\theta)}{M_0} - 1 \right) 10^3 \quad (3.6)$$

then, from Equation 3.5, we have

$$r(\theta-1) = r(\theta) - 1 - \tau \left\{ \ln(\Delta(\theta-1)10^{-3} + 1) - \ln(\Delta(\theta)10^{-3} + 1) \right\}.$$

The annual variation of atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio has been measured experimentally in terms of $\Delta^{14}\text{C}$. Jong (1981) and Stuiver and Quay (1981) report that this variation has different long, medium and short-term behaviours. Since the knots of the calibration data are approximately every twenty years we are only interested in the very short-term variation. Damon *et. al.* (1989) performed hyper-fine measurements of such levels and from their analysis we can see that in

the very short-term, $\Delta(\theta)$ varies within the range of ± 3 parts per mil. This makes the term $\Delta(\theta)10^{-3}$ small and thus we can use the approximation $\ln(x+1) \approx x$. Hence

$$r(\theta-1) \approx r(\theta)-1-\tau\{\Delta(\theta-1)-\Delta(\theta)\}10^{-3}. \quad (3.7)$$

Letting

$$\alpha(\theta) = \{\Delta(\theta-1)-\Delta(\theta)\}10^{-3}, \quad (3.8)$$

we obtain

$$r(\theta-1) \approx r(\theta)-1-\tau\alpha(\theta). \quad (3.9)$$

This means that $r(\theta-1)$ is equal to $r(\theta)$ minus one year minus the distortion term which now is represented by $\tau\alpha(\theta)$.

We need to make some modelling assumptions about $\alpha(\theta)$ to continue with our analysis. We expect the process $\alpha(\theta)$ to be influenced by many factors such as cosmic radiation, sunspots, geomagnetism, etc. and hence its behaviour will be complex. However, since the atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio varies around an 'equilibrium level' we would expect $E[\alpha(\theta)] = 0$. Furthermore, we will assume that $\alpha(\theta)$ is stationary and that $V[\alpha(\theta)] = s^2$.

As a first approximation we are going to suppose that $\alpha(\theta)$ is a Gaussian white noise process. That is

$$\alpha(\theta) \sim N(0, s^2)$$

where the $\alpha(\theta)$'s are independent. We appreciate that it is quite likely that $\alpha(\theta)$ will have some correlation structure and thus the 'white noise' model for it may not be a good choice. Nevertheless, the 'white noise' model will facilitate our analysis and will also give us results that can be generalised for more complicated

model choices for $\alpha(\theta)$. In the next Section we will study the adequacy of the 'white noise' model and also consider an alternative model for $\alpha(\theta)$ with a correlation structure.

It is simple to see that given the 'white noise' model for $\alpha(\theta)$, $r(\theta)$ will have a normal distribution for every θ . Thus we only need now to calculate $E[r(\theta)]$ and $V[r(\theta)]$ to have a complete specification for $r(\theta)$. To do this we use the auxiliary process $X(\theta) = r(\theta) - \theta$ and Equation 3.9 becomes

$$X(\theta-1) = X(\theta) - \tau\alpha(\theta).$$

From Equation 3.2 we see that, given the calibration data z , $X(\theta)$ is equal to $(X(\theta) \mid z_k, z_{k-1})$ for $t_k > \theta \geq t_{k-1}$. This means that we only need to know the distribution for $X(\theta)$ given z_k and z_{k-1} to have a complete definition for the distribution of $X(\theta)$ (and thus for $r(\theta)$).

Given z_k it is easy to see from the above formula that

$$X(t_k - \theta) = X(t_k) - \tau\{\alpha(t_k - 1) + \alpha(t_k - 2) + \dots + \alpha(t_k - \theta)\}$$

for integer $\theta > 0$. Now, letting $\eta(i-1) = -\tau\alpha(t_k - i)$, we have

$$X(t_k - \theta) = X(t_k) + \eta(0) + \eta(1) + \dots + \eta(\theta - 1).$$

(Note that $\eta(i)$ has mean 0 and standard deviation $\sqrt{V[\eta(i)]} = s\tau$ calendar years.)

Letting

$$\gamma(\theta) = \eta(0) + \eta(1) + \dots + \eta([\theta] - 1)$$

we have

$$X(t_k - \theta) = X(t_k) + \gamma(\theta) \tag{3.10}$$

where $\gamma(0)$ is defined to be 0. Notice from Equation 3.1 that $X(t_k) =$

$N(x_k - t_k, \sigma_k^2)$. Assuming that $X(t_k)$ and $\gamma(\theta)$ are independent then we only need to calculate the distribution of $\gamma(\theta)$ to know the distribution of $X(t_k - \theta)$. We will do this by identifying $\gamma(\theta)$ with a 'Wiener process' (if we rescale the $\eta(i)$'s to have variance 1). That is, letting $l = t_k - t_{k-1}$, we see from Billingsley (1968, p.62-65) that $\frac{1}{s\tau\sqrt{l}}\gamma(\theta)$ is the sum of random variables that converges in distribution to the Wiener process $W(t)$, with $t = \frac{\theta}{l}$, when l , the number of elements in the series, is large.

Given now both z_k and z_{k-1} we have $X(t_k - l) = X(t_{k-1})$ and thus we should have $\gamma(l) = X(t_{k-1}) - X(t_k)$. From Billingsley (1968, p.62-65) we see that

$$\frac{1}{s\tau\sqrt{l}} \left\{ \gamma(\theta) - \{X(t_{k-1}) - X(t_k)\} \left(\frac{\theta}{l} \right) \right\}$$

is the sum of random variables that converges in distribution to the 'Brownian Bridge' $W^0(t)$ with $t = \frac{\theta}{l}$. Thus, conditional upon both z_k and z_{k-1} , $X(t_k - \theta)$ is approximately equal in distribution to

$$X(t_k) + \{X(t_{k-1}) - X(t_k)\} \left(\frac{\theta}{l} \right) + W^0 \left(\frac{\theta}{l} \right) s\tau\sqrt{l}.$$

Since $E \left[W^0 \left(\frac{\theta}{l} \right) \right] = 0$, it is simple to see that (using Equation 3.1)

$$E[X(t_k - \theta)] = \left(1 - \frac{\theta}{l} \right) (x_k - t_k) + \left(\frac{\theta}{l} \right) (x_{k-1} - t_{k-1}),$$

and, since $E[X(t_k - \theta)] = E[r(t_k - \theta)] - (t_k - \theta)$, the above expression implies that

$$E[r(\theta) \mid z_k, z_{k-1}] = x_k \left(\frac{\theta - t_{k-1}}{t_k - t_{k-1}} \right) + x_{k-1} \left(\frac{t_k - \theta}{t_k - t_{k-1}} \right) = \mu(\theta);$$

$$t_k > \theta \geq t_{k-1}, k = 1, 2, \dots, n.$$

where $t_0 = 0$ and $x_0 = 0$.

In other words, the expectation of the calibration curve is the piece-wise linear calibration curve $\mu(\theta)$, as defined in Equation 1.3. This was to be

anticipated since $E[r(t_k) | z_k, z_{k-1}] = x_k = \mu(t_k)$ and thus the above expectation is simply the linear interpolation of this property between the knots. Since $r(\theta)$ is normally distributed we only need now to calculate $V[r(\theta)]$ to have a complete specification for $r(\theta)$. This is not difficult to do since

$$V\left[W^0\left(\frac{\theta}{l}\right)\right] = \frac{\theta}{l}\left(1 - \frac{\theta}{l}\right), \quad (3.11)$$

and from this we see that

$$V[X(t_k - \theta)] \approx \left(1 - \frac{\theta}{l}\right)^2 \sigma_k^2 + \left(\frac{\theta}{l}\right)^2 \sigma_{k-1}^2 + \frac{(s\tau)^2}{l} \theta(l - \theta).$$

The above formula can then be rearranged to give the following expression

$$V[X(t_k - \theta)] \approx \frac{1}{l^2} ((l - \theta)^2 \sigma_k^2 + \theta^2 \sigma_{k-1}^2) + \frac{(s\tau)^2}{l} \theta(l - \theta)$$

for $0 \leq \theta \leq l$, since $V[X(t_k)] = \sigma_k^2$ (see Equation 3.1). Remembering that $V[X(t_k - \theta)] = V[r(t_k - \theta)]$ and that $l = t_k - t_{k-1}$ we have

$$V[r(\theta)] = \left(\frac{\theta - t_{k-1}}{t_k - t_{k-1}}\right)^2 \sigma_k^2 + \left(\frac{t_k - \theta}{t_k - t_{k-1}}\right)^2 \sigma_{k-1}^2 + \frac{(s\tau)^2 (\theta - t_{k-1})(t_k - \theta)}{(t_k - t_{k-1})};$$

$$t_k > \theta \geq t_{k-1}, k = 1, 2, \dots, n \quad (3.12)$$

where $t_0 = 0$ and $\sigma_0 = 0$.

It is important at this point to stop and summarise what we have done so far. Given Equation 3.1 and Equation 3.2, and the assumption that $\alpha(\theta)$ is a Gaussian white noise process, we have extended the definition of $r(\theta)$ to the entire calendar scale. That is, we now know that

$$r(\theta) \sim N(\mu(\theta), \sigma^2(\theta)) \quad (3.13)$$

where $\sigma^2(\theta) = V[r(\theta)]$ and $\mu(\theta)$ is the piece-wise linear calibration curve. This means that $r(\theta)$ is now modelled as a calibration curve that varies around $\mu(\theta)$ (the piece-wise linear calibration curve) with variance $\sigma^2(\theta)$. Notice that at each

knot, t_k , the variance corresponds to the variance reported in the high-precision calibration data for that knot, σ_k^2 . Away from the knots, $\sigma^2(\theta)$ is a piece-wise quadratic function of θ that depends on the variances at the nearest knots, and on s , the variance of the white noise process $\alpha(\theta)$. Note that $\sigma^2(\theta)$ increases as we move away from one knot and decreases when approaching the next (see Figure 3.2, in the next Section). This is to be expected since we have only observed $r(\theta)$ at the knots. Consequently, the variance must be *smaller* where we have observed $r(\theta)$ than where we have not.

Later in Section 3.2 we develop a new calibration method for radiocarbon determinations based upon the process $r(\theta)$. This new calibration method will consider the variances σ_k^2 's in the high-precision calibration data, now included through the variance of $r(\theta)$, that were neglected in the calibration method used in Chapter 1. Note that to do so we need to estimate s . In the next Section this will be carried out using some published data.

3.1.2 Estimation of s

The atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio is influenced by several factors and constitutes a very complex process which can vary greatly over a period of hundred years. However, within the range of one to twenty years, the atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio is reasonable stable (Stuiver and Quay 1981) and has been estimated by measuring the $^{14}\text{C}/^{12}\text{C}$ ratio in tree-rings. This is done in terms of $\Delta(\theta)$ (see Equation 3.6) and from that we calculate $\alpha(\theta)$ using Equation 3.8. Now, only short-lived trees have rings wide enough and so suitable for a radiocarbon analysis. Therefore it is only possible to obtain values of $\alpha(\theta)$ for the last hundred years. In fact, at the present time it is difficult to detect the *natural* variations of atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio because of the radioactive contamination arising from nuclear testing that has occurred since 1950. Thus, measurements of

$\alpha(\theta)$ are possibly available only for the period 1900 to 1950. Measurements for $\Delta(\theta)$, the atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio in year θ , are given by Damon *et. al.* (1989), who made such measurements on an annual basis for the period 20-0 BP (1930-50 AD). A plot showing the values of $\alpha(\theta)$ can be seen in Figure 3.1.

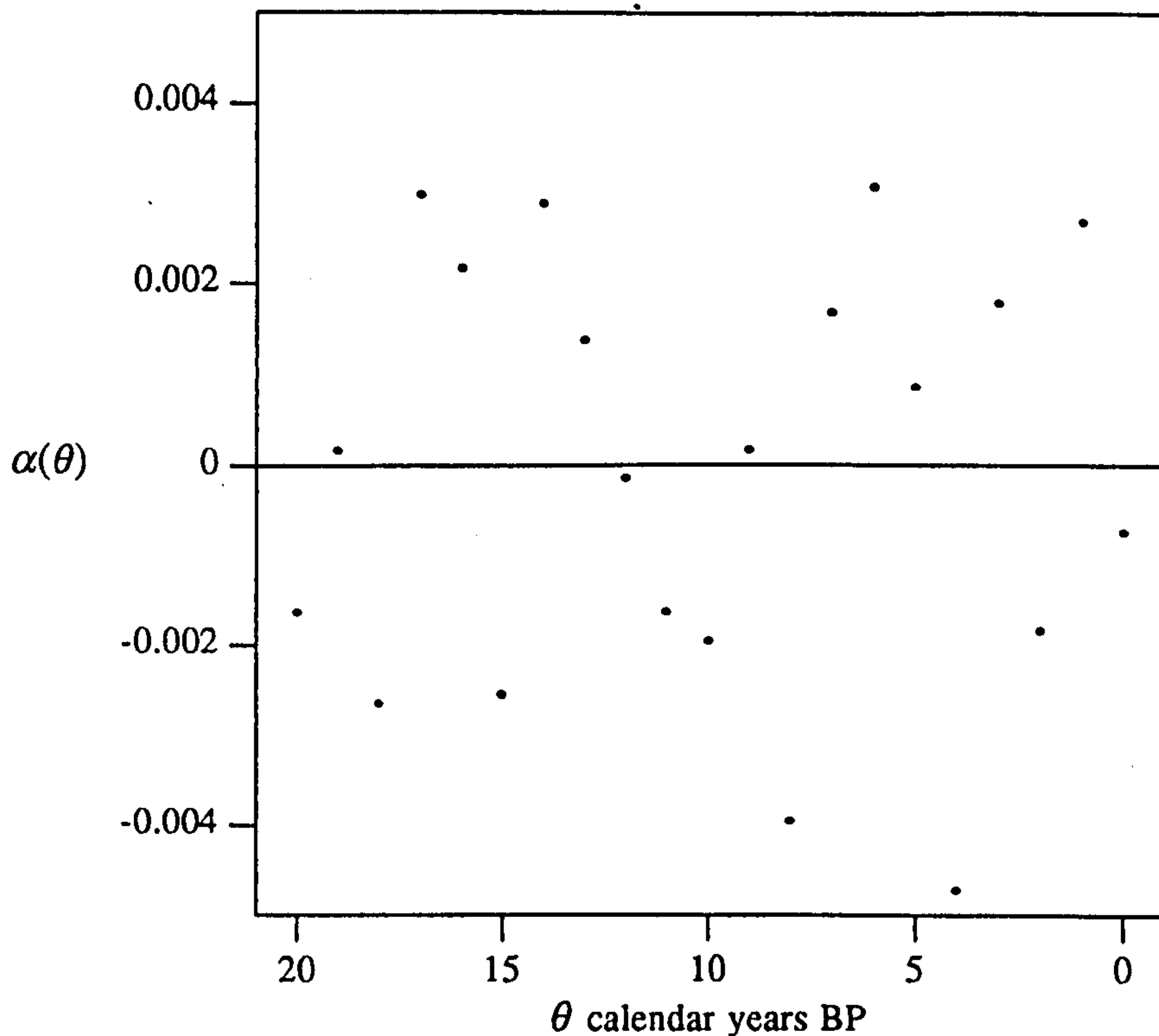


Figure 3.1
Plot of values of $\alpha(\theta)$ taken from Damon *et. al.* (1989),
for the units of $\alpha(\theta)$ see Equations 3.6 and 3.8.
(0 BP is equivalent to 1950 AD.)

The sample standard deviation of $\alpha(\theta)$ is 0.00235. If we use this as an approximation for the standard deviation, s , of $\alpha(\theta)$ we have

$$\alpha(\theta) \sim N(0, (0.00235)^2).$$

Since $\tau = 8033$ (the radiocarbon mean life) we have $s\tau \approx 19$. Using this we can calculate $\sigma^2(\theta)$ in Equation 3.13, which is the estimated variance for the

calibration curve at any point.

We know that $\sigma^2(t_k) = \sigma_k^2$, the variance of the calibration curve at a knot is the variance reported in the high-precision data for that knot. Between knots, this variance will increase and decrease to match the variance for the next knot, following a 'Brownian Bridge like' process. To illustrate this phenomenon we present in Figure 3.2 a section of the piece-wise linear calibration curve $\mu(\theta)$ (which is the expectation of $r(\theta)$) and the curves $\mu(\theta) + \sigma(\theta)$ and $\mu(\theta) - \sigma(\theta)$. The process then varies around $\mu(\theta)$ with standard deviation $\sigma(\theta)$.

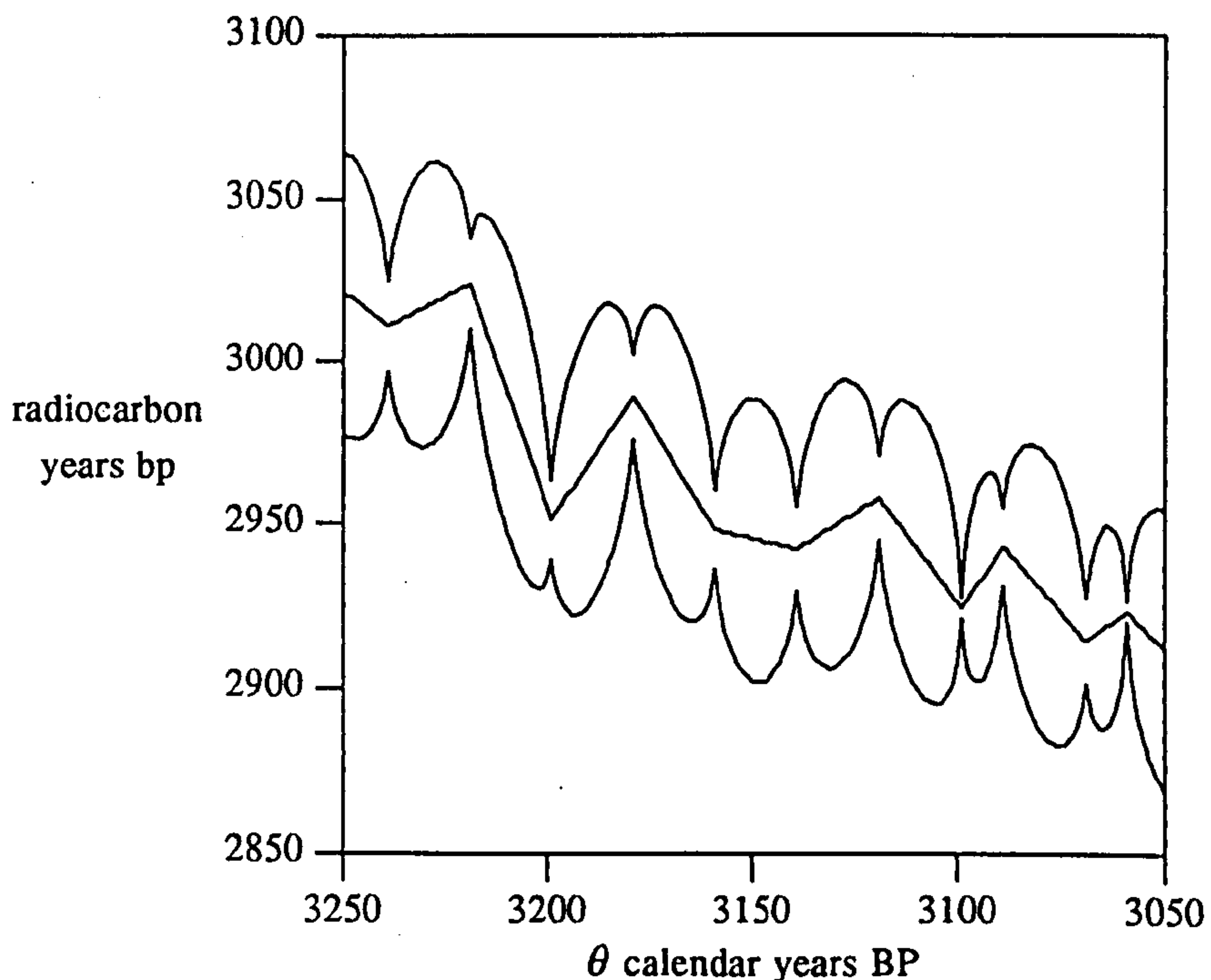


Figure 3.2

Section of the piece-wise linear calibration curve $\mu(\theta)$ and $\mu(\theta) \pm \sigma(\theta)$ ($s\tau = 19$).
The calibration curve $r(\theta)$ varies around $\mu(\theta)$ with variance $\sigma^2(\theta)$.

From Figure 3.2 we notice that the standard deviation $\sigma(\theta)$ seems large in comparison to the variances at the knots. This may be a consequence of two factors. These are

(i) $\alpha(\theta)$ cannot be represented by a 'white noise' process

and

(ii) overestimating s .

Both (i) and (ii) seem possible. As we only have very few measurements for $\alpha(\theta)$, then it is likely that we could be overestimating s , if we use an estimate for the sample variance. Of even bigger concern is the independence of the $\alpha(\theta)$'s. To study this we plotted the first values of the observed autocorrelation function ρ_k in Figure 3.3.

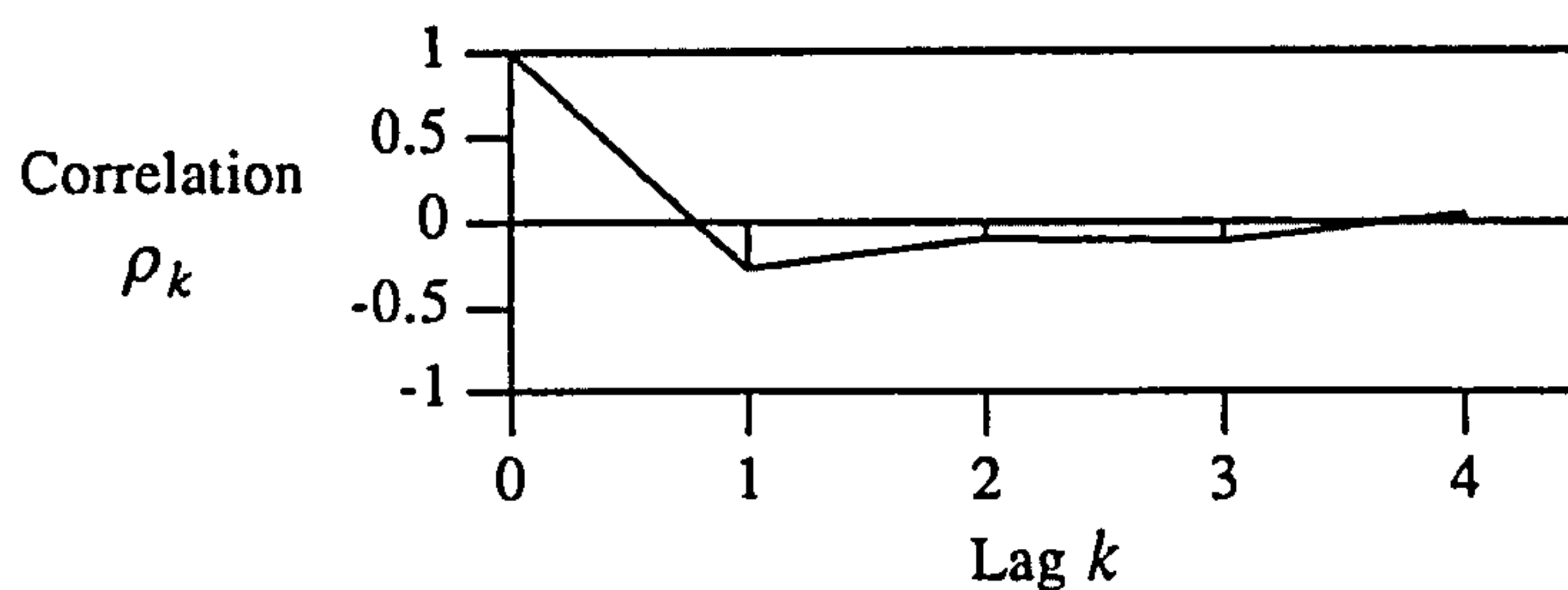


Figure 3.3
First five values of the sample autocorrelation function for the process $\alpha(\theta)$.

From Figure 3.3 we see that some correlation structure in $\alpha(\theta)$ seems possible. Moreover recalling the definition of $\alpha(\theta)$ we see that

$$\alpha(\theta) = \{\Delta(\theta) - \Delta(\theta-1)\} 10^{-3},$$

which is the first difference of the annual atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio $\Delta(\theta)$. Indeed, we might expect the annual atmospheric $^{14}\text{C}/^{12}\text{C}$ ratios to be correlated. In particular we expect that $\Delta(\theta)$ may be dependent upon the atmospheric $^{14}\text{C}/^{12}\text{C}$ ratios of the previous few years, and thus an $\text{AR}(p)$ model for $\Delta(\theta)$ seems reasonable. That is

$$\Delta(\theta) = \sum_{j=1}^p \phi_j \Delta(\theta-j) + z(\theta),$$

where the ϕ_j 's are constants and $z(\theta)$ is Gaussian white noise. Taking this as a model for $\Delta(\theta)$, it is easy to see that $\alpha(\theta)$ will be an ARMA($p, 2$) model. It will be very difficult to obtain reliable estimates for the parameters ϕ_j 's since we have very little data for $\Delta(\theta)$. Fortunately, as we will see, for the purposes of estimating the variance of the calibration curve $r(\theta)$, this will not be necessary.

Given the above considerations it seems reasonable to suppose that $\alpha(\theta)$ is an ARMA(p, m) process. Assuming $\alpha(\theta)$ to be stationary we see that (Box and Jenkins 1976, p. 79)

$$\sum_{k=1}^{\infty} \rho_k^2 < \infty.$$

That is, $\alpha(\theta)$ has a correlation structure that damps to zero as the 'lag' k increases.

Using this it is easy to see from Billingsley (1968, p. 174) that

$$V\left[W^0\left(\frac{\theta}{l}\right)\right] = \frac{\theta}{l}\left(1 - \frac{\theta}{l}\right)(1 + 2\rho^*),$$

where $\rho^* = \sum_{k=1}^{\infty} \rho_k$ (see Equation 3.11). Here the term ρ^* accounts for the correlation structure in $\alpha(\theta)$. Hence using Equation 3.12 we have

$$\sigma^2(\theta) = \left(\frac{\theta - t_{k-1}}{t_k - t_{k-1}}\right)^2 \sigma_k^2 + \left(\frac{t_k - \theta}{t_k - t_{k-1}}\right)^2 \sigma_{k-1}^2 + \frac{\lambda^2(\theta - t_{k-1})(t_k - \theta)}{(t_k - t_{k-1})};$$

$$t_k > \theta \geq t_{k-1}, k = 1, 2, \dots, n. \quad (3.14)$$

where $t_0 = 0$, $\sigma_0 = 0$ and $\lambda = s\tau\sqrt{1 + 2\rho^*}$.

It is difficult to estimate ρ^* because of the small amount of data available for $\alpha(\theta)$. In fact, none of the ρ_k 's can be estimated with any reasonable reliability. However, since we expect the atmospheric $^{14}\text{C}/^{12}\text{C}$ ratios to depend only on the

previous few years, it is reasonable to expect that only the first one or two terms in $\sum_{k=1}^{\infty} \rho_k$ will be significant. The first few sample autocorrelations are -0.27, -0.08, -0.10, 0.04 and thus we would expect $\rho^* \approx -0.4$. This will give us $\lambda = s\tau\sqrt{1+2\rho^*} \approx 8.5$, resulting in a smaller value for $\sigma^2(\theta)$ between the knots. We present in Figure 3.4 a section (same as in Figure 3.2) of the piece-wise linear calibration curve $\mu(\theta)$ and the curves $\mu(\theta) + \sigma(\theta)$ and $\mu(\theta) - \sigma(\theta)$ based on the latest model for $\alpha(\theta)$.

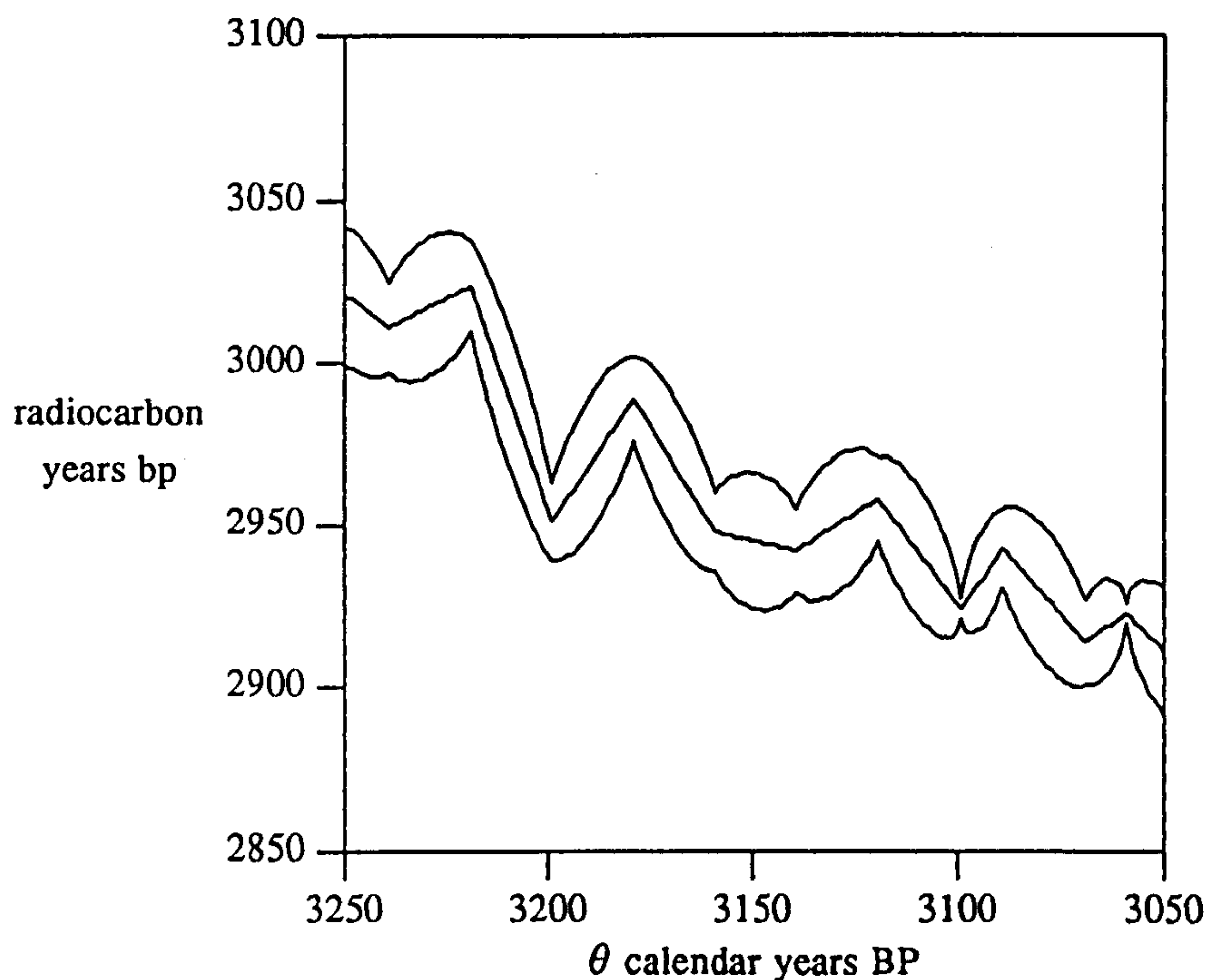


Figure 3.4
Section of the piece-wise linear calibration curve $\mu(\theta)$ and $\mu(\theta) \pm \sigma(\theta)$
with $s\tau\sqrt{1+2\rho^*} = 8.5$.
The calibration curve $r(\theta)$ varies around $\mu(\theta)$ with variance $\sigma^2(\theta)$.

From the above discussion we observe that λ in Equation 3.14 is in the range of 8.5 to 19 (radiocarbon years) and we expect 19 to be an overestimate. In the next Section we develop a new calibration procedure for radiocarbon determinations that uses the variance for the calibration curve $\sigma^2(\theta)$. Then in Section 3.3 we compare the resulting posterior distributions for the calendar dates of some

radiocarbon determinations using this new calibration method and the one presented in Chapter 1. We prefer to be 'pessimistic' and will use the estimate of $\lambda = 19$ to perform the calculations. Even so, the differences between the results obtained using the calibration method given here and the one presented in Chapter 1 are, apart from some exceptions, negligible.

3.2 Calibration

Within the Bayesian framework there is an established method for calibration problems when we have random observations as with the Pearson and Stuiver data (or even with random knots, which is not our case). A standard reference for this subject is Aitchison and Dunsmore (1975, chapter 10). Suppose we want now to calibrate a new radiocarbon determination $y \pm \sigma$. Given its 'true' radiocarbon age t , we have

$$y \mid t, \sigma^2 \sim N(t, \sigma^2).$$

Again for simplicity we avoid explicitly conditioning on σ^2 and simply write $y \mid t$. Let θ be the associated calendar year of determination $y \pm \sigma$. It is easy to see from Section 3.1 that given a particular calendar year θ , its corresponding radiocarbon age is equal to $r(\theta)$, which is normally distributed with mean $\mu(\theta)$ and variance $\sigma^2(\theta)$. Therefore we have

$$t \mid \theta \sim N(\mu(\theta), \sigma^2(\theta)).$$

The process of calibration consists of finding $f(y \mid \theta)$ from the distributions of $f(y \mid t, \theta)$ and $f(t \mid \theta)$, and then using $f(y \mid \theta)$ for our likelihood (the 'calibrated likelihood'). It is easy to see that

$$f(y \mid \theta) = \int f(y \mid t, \theta) f(t \mid \theta) dt.$$

Firstly, to evaluate the integral, we note that given both t and θ , y only depends on t (the radiocarbon age), and this means that $f(y | t, \theta) = f(y | t)$. Secondly, we note that both functions in the integral represent normal densities, and using a standard result, we have

$$y | \theta \sim N(\mu(\theta), \sigma^2(\theta) + \sigma^2).$$

Thus our likelihood is again normal as in Section 1.2, but now corrected with the variance $\sigma^2(\theta)$, reflecting the errors in the calibration data. Using a vague prior distribution for θ , we have that the posterior distribution for θ given y is

$$f(\theta | y) \propto \frac{1}{\sqrt{\sigma^2(\theta) + \sigma^2}} \exp \left\{ -\frac{1}{2} \left(\frac{(y - \mu(\theta))^2}{\sigma^2(\theta) + \sigma^2} \right) \right\}.$$

To facilitate the notation we let $\omega^2(\theta) = \sigma^2(\theta) + \sigma^2$ and thus $y | \theta \sim N(\mu(\theta), \omega^2(\theta))$ and

$$f(\theta | y) \propto \frac{1}{\omega(\theta)} \exp \left\{ -\frac{1}{2} \left(\frac{(y - \mu(\theta))^2}{\omega^2(\theta)} \right) \right\}. \quad (3.15)$$

In subsequent Chapters we will work with sets, or groups of sets, of radiocarbon determinations represented by $y_j \pm \sigma_j$'s or $y_{i,j} \pm \sigma_{i,j}$'s. We will use the functions

$$\omega_j^2(\theta_j) = \sigma^2(\theta_j) + \sigma_j^2 \quad \text{and} \quad \omega_{i,j}^2(\theta_{i,j}) = \sigma^2(\theta_{i,j}) + \sigma_{i,j}^2, \quad (3.16)$$

to denote the overall variances. This notation should cause no confusion.

3.3 Examples

Following the considerations of Section 3.1.2, we calculate $\sigma^2(\theta)$ and the resulting $f(\theta | y)$ from Equation 3.15 using numerical integration. Examples of these distributions can be found in Figures 3.5 and 3.6. These arise from the same determinations as for the distributions in Figures 1.3 and 1.4 (which were

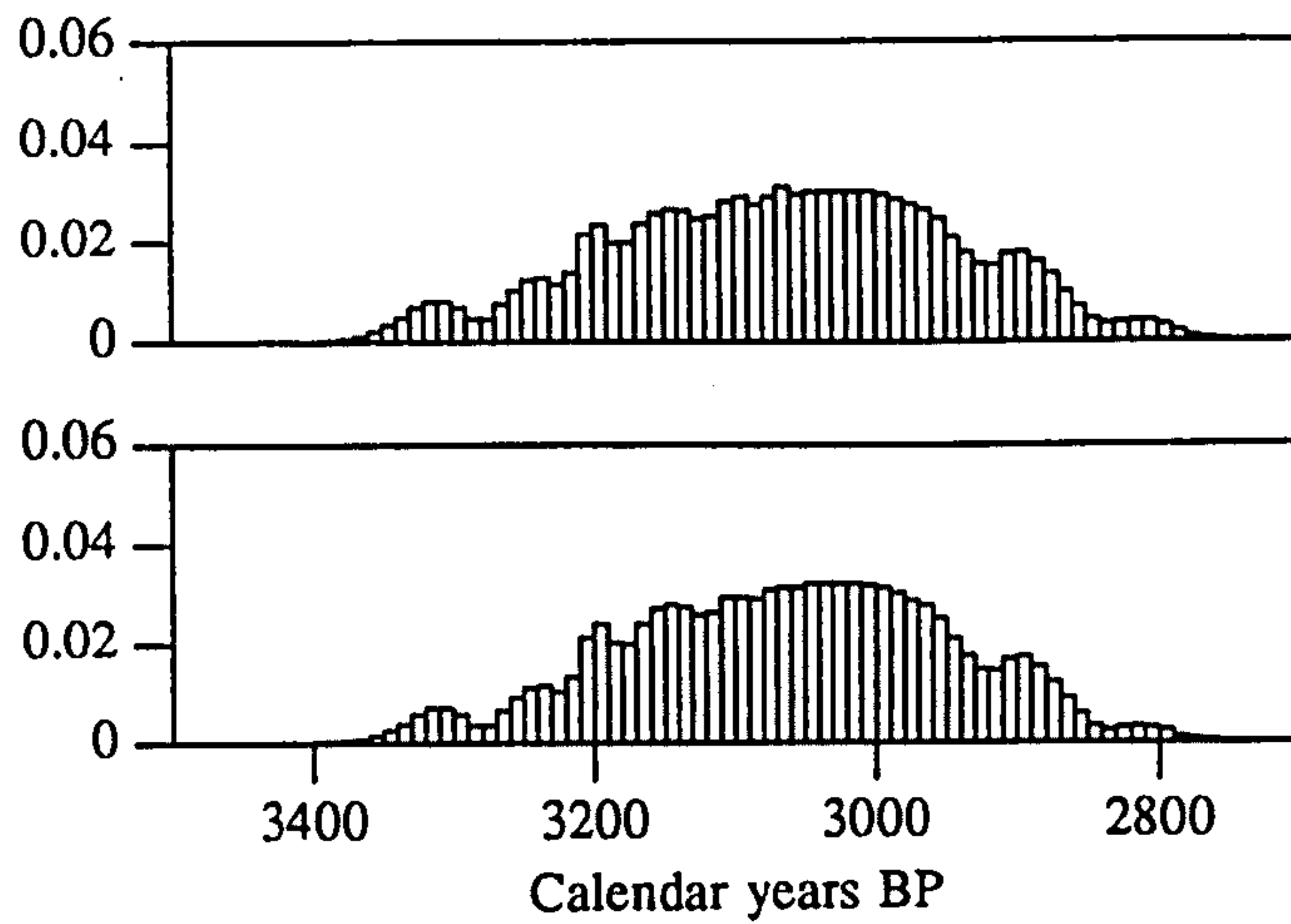


Figure 3.5

Histograms showing the posterior distribution for radiocarbon determination 2900 ± 80 , considering the errors (above) and not considering the errors in the calibration data (below).

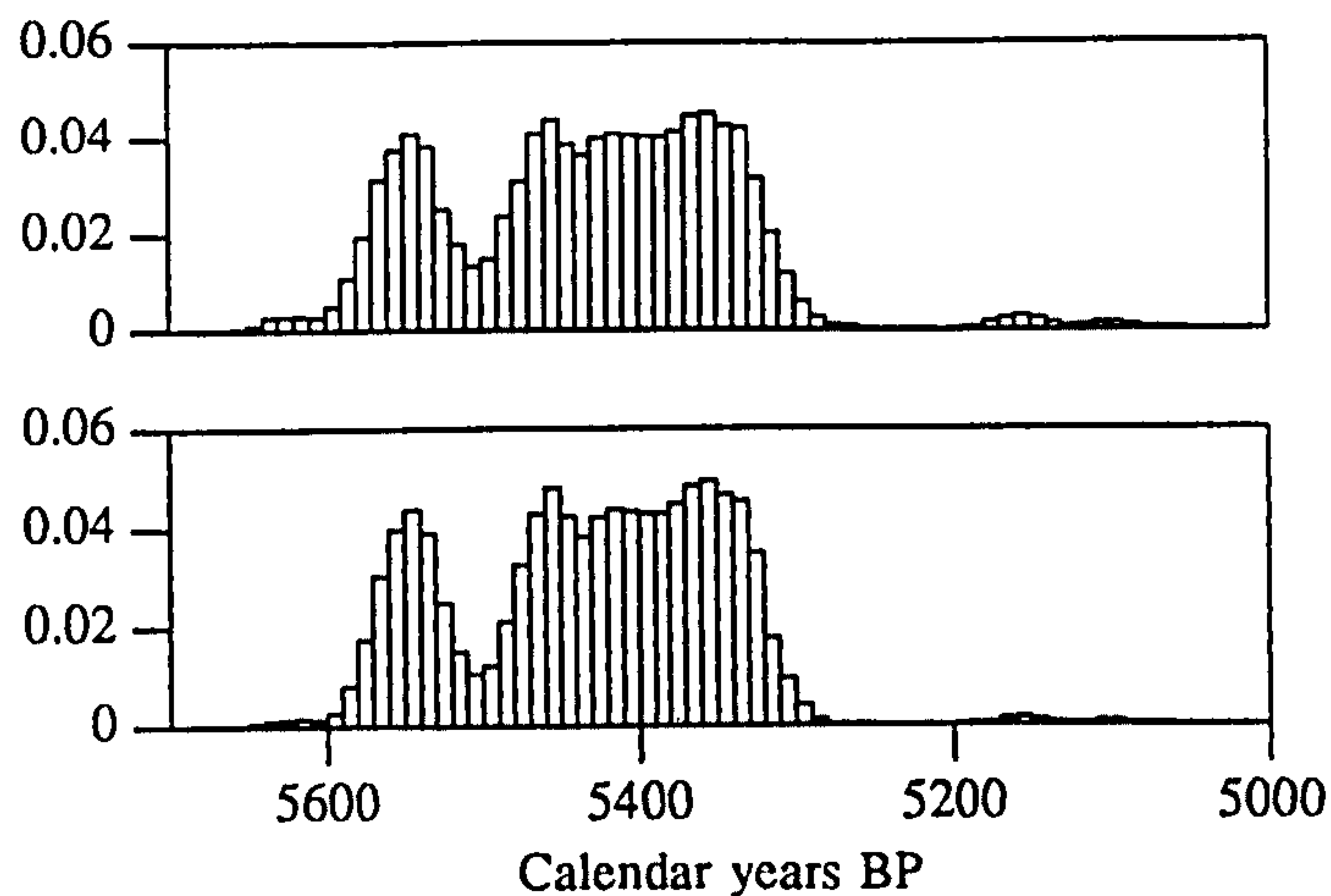


Figure 3.6

Histograms showing the posterior distribution for radiocarbon determination 4700 ± 60 , considering the errors (above) and not considering the errors in the calibration data (below).

calculated with the conventional calibration procedure given in Chapter 1).

Comparing the distributions in Figure 3.5 and in Figure 3.6 we see that there is little difference between using the conventional calibration technique presented in Section 1.2 and the alternative technique presented here. From the point of

view of archaeological interpretation, there is virtually no difference. The standard deviations for the radiocarbon determinations in this case are 80 and 60 (whereas the standard deviations for the calibration data, σ_k 's, for this section of the calendar scale are between 15 to 20 years).

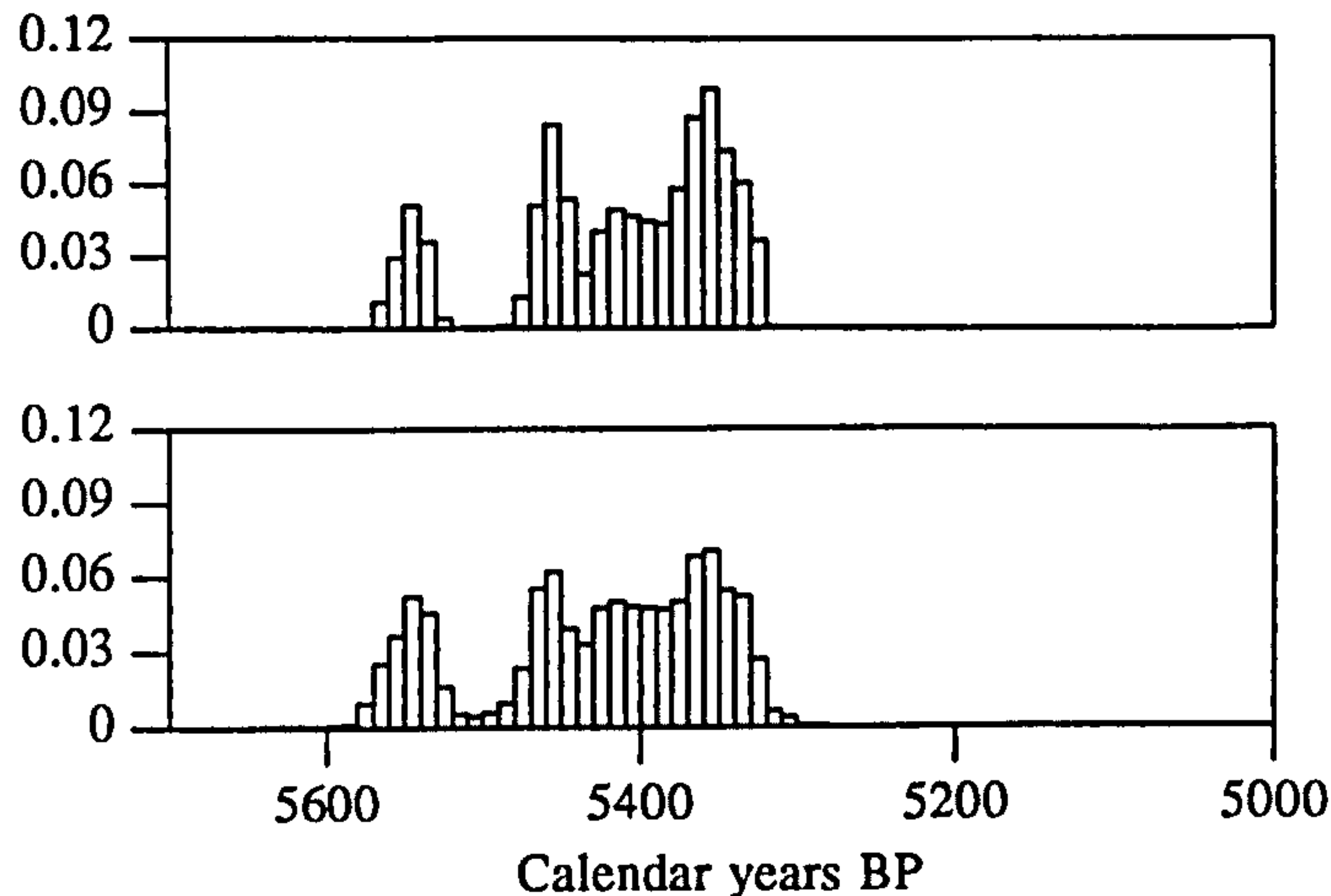


Figure 3.7

Histogram showing posterior distribution for radiocarbon determination 4700 ± 25 , not considering the errors (above) and considering the errors in the calibration data (below).

If we now reduce the standard deviation we see that differences start to appear. In Figure 3.7 we have a comparison of the distributions with standard deviation of 25, and in Figures 3.8 and 3.9 for standard deviations of 15 and 5 respectively. These standard deviations correspond to high-precision radiocarbon determinations, or alternatively an averaged determination derived from several 'replications'.

From Figures 3.5 to 3.9 we note that for a standard deviation of 5 there are substantial differences in the resulting posterior distributions depending on the calibration technique used. Such differences become less evident as we have higher standard deviations, and even with a standard deviation as low as 25 years, the difference could possibly be neglected. In general, any difference will depend

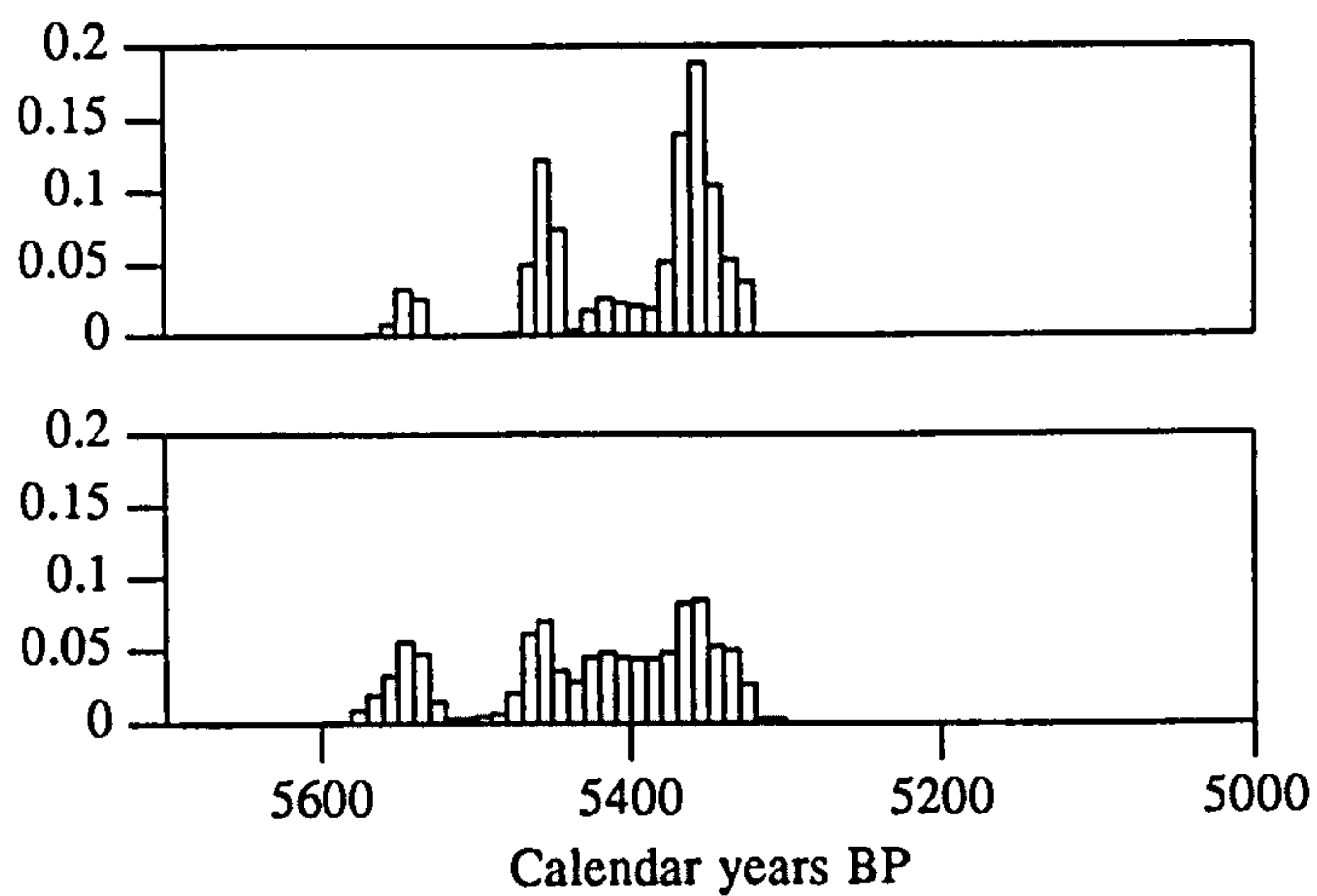


Figure 3.8

Histogram showing posterior distribution for radiocarbon determination 4700 ± 15 , not considering the errors (above) and considering the errors in the calibration data (below).

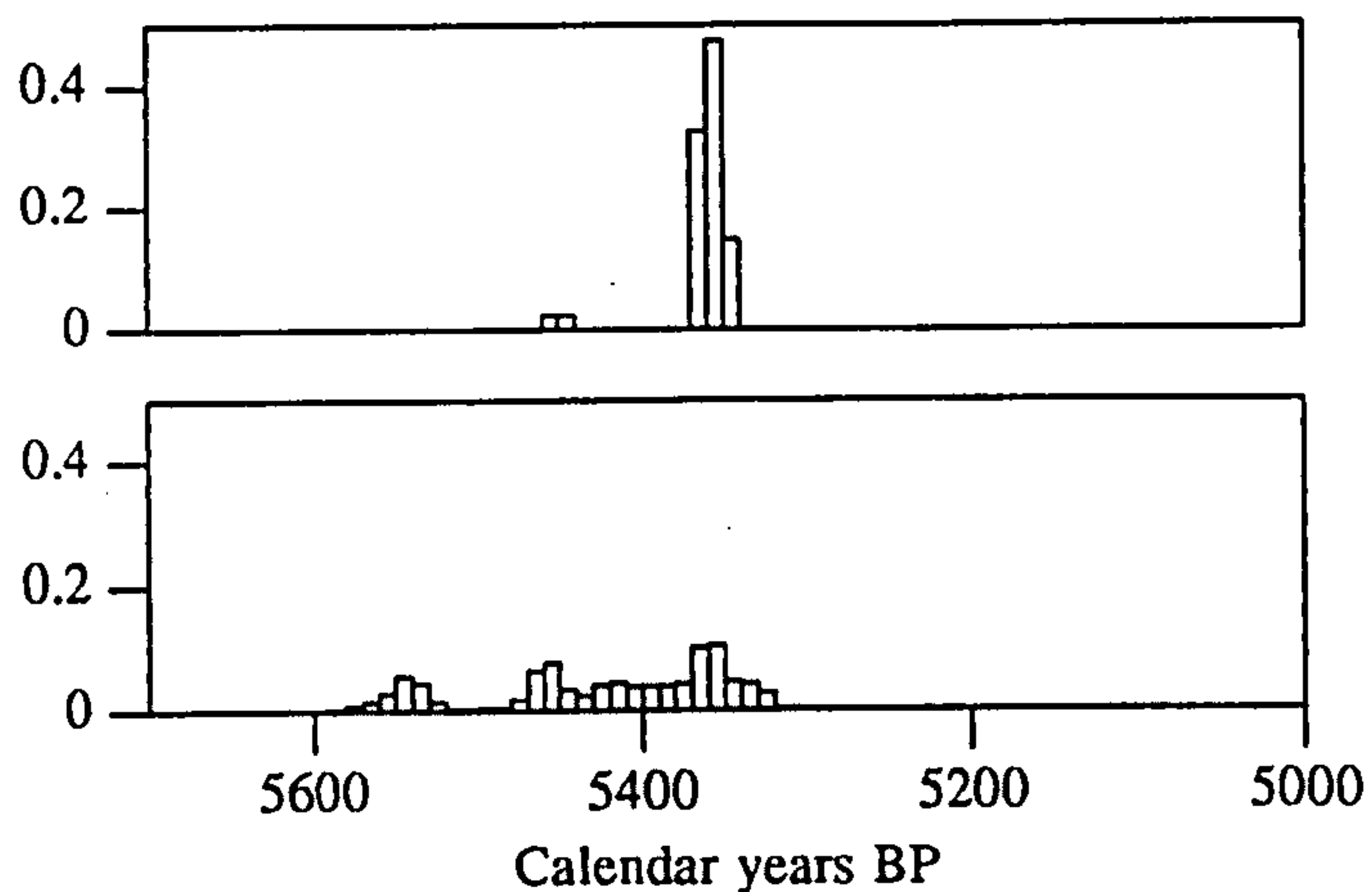


Figure 3.9

Histogram showing posterior distribution for radiocarbon determination 4700 ± 5 , not considering the errors (above) and considering the errors in the calibration data (below).

upon the section of the calibration curve we are working with. However, it seems that only if we are dealing with very low standard deviations (of 25 to 15 years or less) we will find a difference which could be of archaeological significance. In such circumstances, the errors in the calibration curve should be incorporated into the calibration method employed. Otherwise we could be working with artificially

precise calibrated distributions.

Given the complicated modelling and the simulation techniques to be used in Chapters 4, 5, 6 and 7, a fast and simple calibration process will be much appreciated. The calibration process explained here does involve several more calculations than the conventional one given in Chapter 1 and, in some situations, may represent an important extra computational cost. For practical reasons, in some of the examples presented in the next Chapters we have not used the calibration process developed in this Chapter since the standard deviations of the samples are larger than 30 or 40 years. However, in the majority of the examples we have used our novel calibration process.

3.4 Comparison with other estimates for the variance in $r(\theta)$

In recent developments concerning the calibration of radiocarbon determinations, some authors have proposed other estimates for the variance in the calibration curve (see Section 2.3.5). These estimates usually arise from practical considerations and/or intuitive ideas. From these we identify three types of estimates.

- (i) $\sigma_0^2(\theta) = 0$. That is, the variance in the calibration curve is considered small in relation to the variance of a determination to be calibrated and thus may be ignored (we used this estimate in Chapter 1 and is used by Pazdur and Michczynska 1989, Robinson 1986 and van der Plicht 1990, among others, in their calibration methods) .
- (ii) Another estimate is $\sigma_c^2(\theta) = k^2$, where $k = 10$ or 20 , depending on where on the calendar scale we are. This estimate is simply based on the common values for the σ_k^2 's which are precisely in the range of 10^2 to 20^2 (Leese 1988 and Weninger 1986 propose this estimate in their calibration methods).

(iii) An estimate for the standard deviation (and thus, by squaring, for the variance) in the calibration curve proposed by Stuiver and Reimer (1986) is the linear interpolation of the standard deviations σ_k 's between consecutive knots. That is

$$\sigma_p(\theta) = \left(\frac{\theta - t_{k-1}}{t_k - t_{k-1}} \right) \sigma_k + \left(\frac{t_k - \theta}{t_k - t_{k-1}} \right) \sigma_{k-1}; t_k > \theta \geq t_{k-1}, k = 1, 2, \dots, n.$$

To these we must add our estimate $\sigma^2(\theta)$ that can be interpreted as a quadratic interpolation of the variances σ_k^2 's (see Equation 3.14). To compare these estimates we plot all of them in Figure 3.10 in a particular section of the calendar scale.

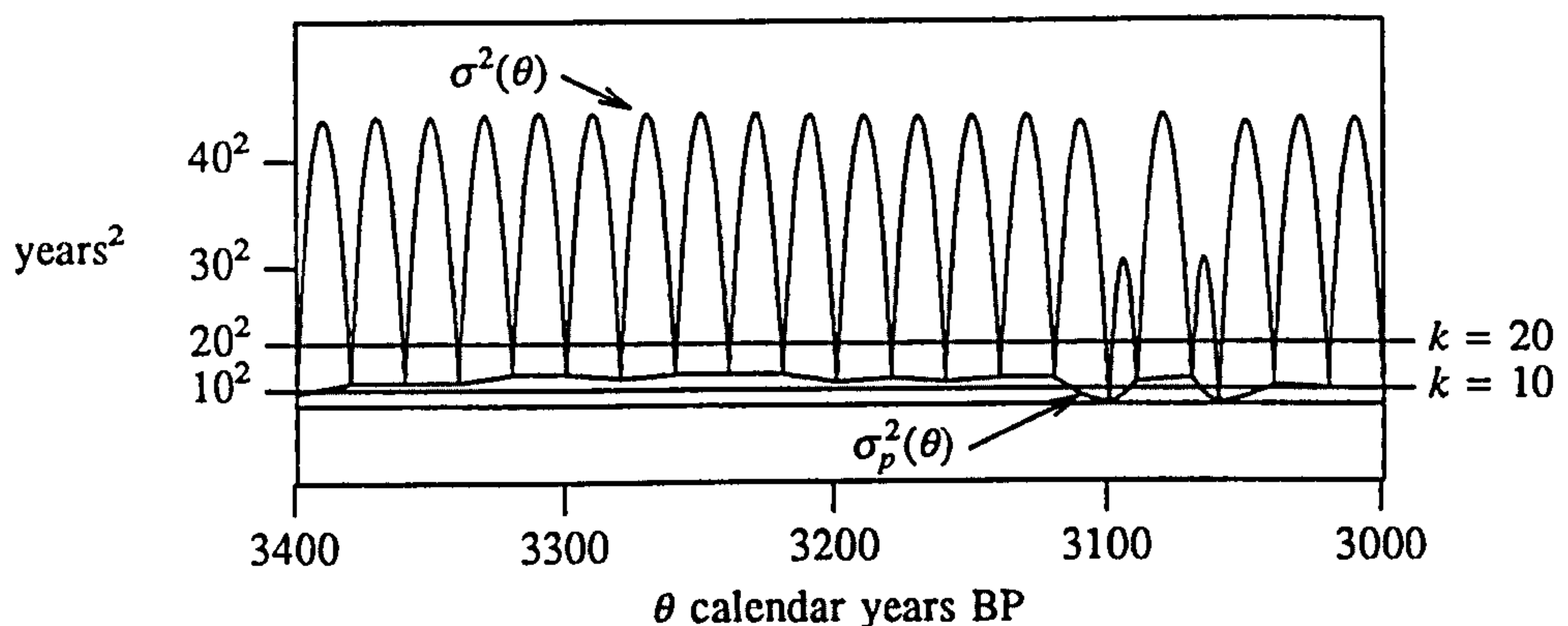


Figure 3.10

Plot comparing various estimates for the variance in the calibration curve, in a particular section of the calendar scale.

As seen in Figure 3.10, the different estimates proposed for the variance in the calibration curve are small in comparison to our estimate of $\sigma^2(\theta)$ (although we must remember that this represents a 'pessimistic' estimate for the variance in $r(\theta)$). Evidently, this will have the consequence that the calibration method developed in this Chapter will produce distributions with the higher spread (variance) followed by those using $\sigma_c^2(\theta)$ or $\sigma_p^2(\theta)$ and those using $\sigma_0^2(\theta)$. To

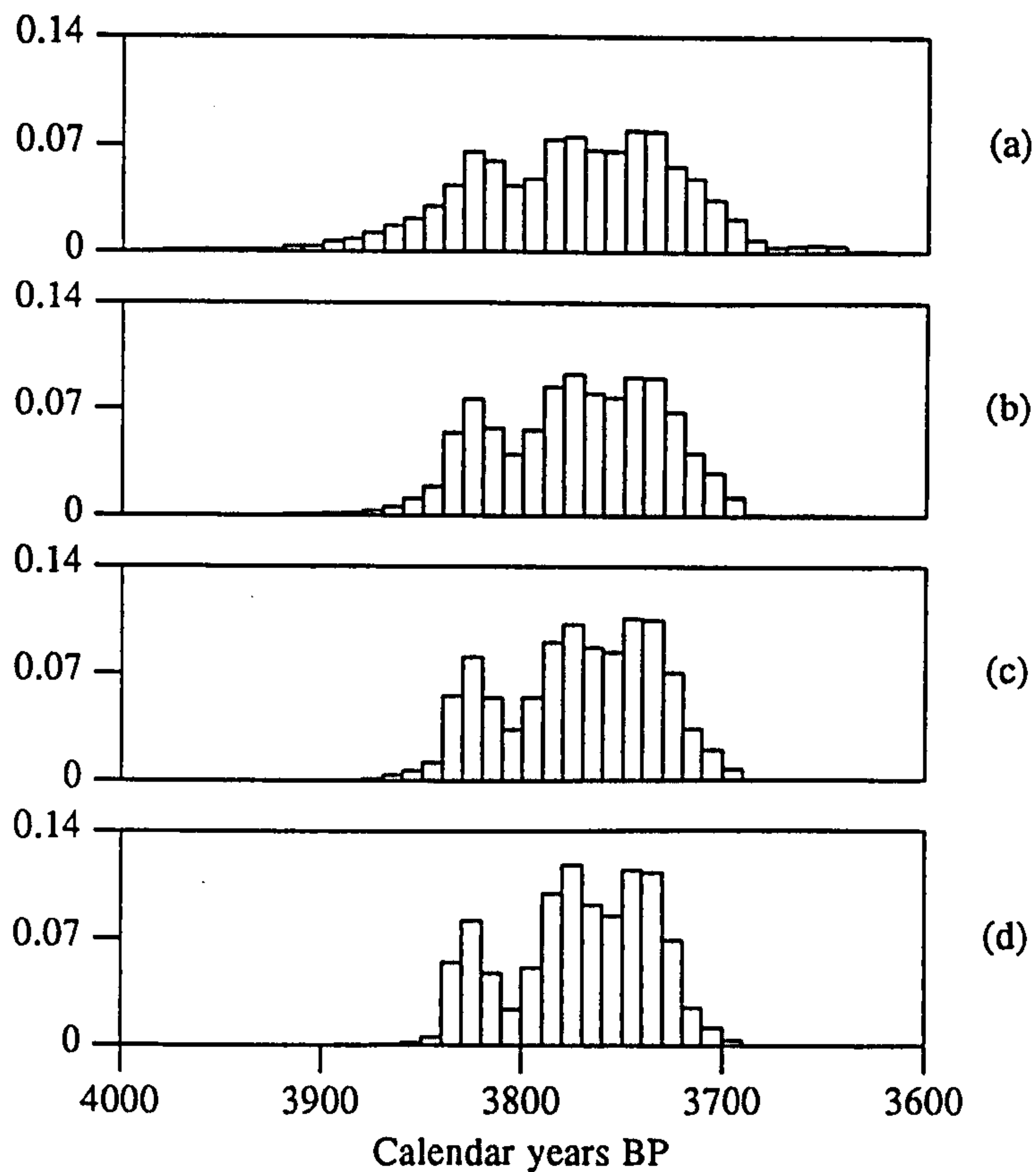


Figure 3.11

Histograms showing posterior distributions for radiocarbon determination 3500 ± 20 ,
 (a) using our method developed in this Chapter (using $\sigma^2(\theta)$),
 (b) using a constant variance in the calibration curve ($\sigma_c^2(\theta) = 20^2$),
 (c) using Stuiver and Reimer's (1986) method (using $\sigma_p^2(\theta)$) and
 (d) ignoring the variance in the calibration curve.

illustrate this in Figure 3.11 we present an example of the resulting distributions when we calibrate the determination 3500 ± 20 using $\sigma^2(\theta)$, $\sigma_c^2(\theta) = 20^2$, $\sigma_p^2(\theta)$ and $\sigma_0^2(\theta)$. Here we use a low standard deviation (20) to appreciate the differences in the calibration methods. The 95% HPD regions for these three distributions are 3870 to 3680 BP, 3840 to 3710 BP, 3830 to 3710 BP and 3830 to 3710 BP, respectively. In this case we see a difference of 30 to 40 years on either side of the HPD regions, between using our calibration method and using $\sigma_c^2(\theta) = 20^2$, $\sigma_p^2(\theta)$ or $\sigma_0^2(\theta)$ (with the latter three methods give effectively the same results).

In the previous Section we have compared the calibration methods using $\sigma^2(\theta)$ and $\sigma_0^2(\theta)$ (histograms (a) and (d) in Figure 3.11). From that comparison we conclude that, only if we are working with low standard errors (25 to 20 years or less) the calibration methods developed here (using $\sigma^2(\theta)$) and the calibration method that ignores the variances σ_k^2 's (using $\sigma_0^2(\theta)$) will produce different distributions, from the point of view of archaeological interpretation. Since our calibration method gives the maximum modification from the resulting distributions where the σ_k^2 's are ignored, therefore all methods (described above) should produce equivalent distributions if we are working with standard deviations of 30 to 40 years or more. Again we stress the fact that if we are working with high-precision determinations, the variance in the calibration curve should be considered in the calibration process to be used.

Chapter 4

Statistical framework

4.1 Introduction

In this Chapter we develop a general framework within which we can tackle the statistical problems associated with the interpretation of radiocarbon determinations. Then in Chapter 5 the archaeological dating problems identified in Chapter 2 will be re-examined in the light of this framework. This will be carried out by considering a series of examples covering a wide variety of problems.

In the earlier Chapters we focused our attention on the radiocarbon calibration technique and some statistical problems related to it. Specifically, we studied the calibration problem and developed our own calibration method in Chapter 3. We believe that the interpretation of radiocarbon determinations involves a wider range of problems than those exclusively related to the radiocarbon dating technique. A radiocarbon determination commonly represents only a part of a complex dating problem in which other (relative or absolute) dating techniques are also involved. In Chapter 1 we have mentioned that the principal aim of radiocarbon dating a series of samples is not just to date the samples themselves but to date the context from which they arise. That is, in most situations, there is an underlying archaeological dating problem that ought to be addressed when interpreting a series of radiocarbon determinations.

We believe that the correct interpretation of radiocarbon determinations and thus, the development of sound statistical techniques for the analysis of radiocarbon determinations, can only be achieved if viewed as part of the problem of archaeological dating in general. This will underpin our approach to developing

our statistical framework. In the next Section we will shift our attention to the archaeological dating problem in general and try to understand the usage of radiocarbon dating from that perspective. This will give us a general outlook on the problem and provide the basis for the framework that will be developed in Section 4.3.

4.2 Dating in archaeology

As we have explained in Chapter 2 and elsewhere in the thesis, radiocarbon dating is a technique for dating objects (as are most science-based dating techniques). Archaeological dating is, however, a much wider problem than simply the dating of individual objects. Only rarely do dates of individual objects allow archaeologists to answer directly the chronological questions they pose.

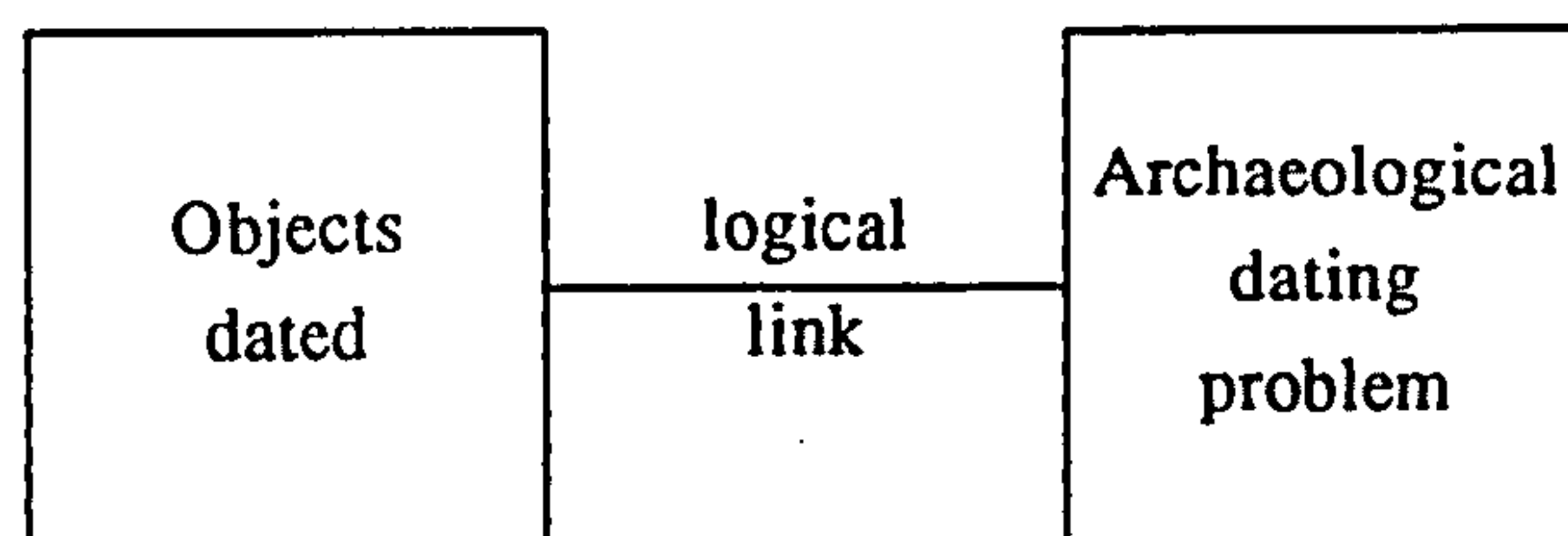


Figure 4.1

A process of logic is needed to link the objects dated with the archaeological dating problem under study. Without such linkage, the dates for the objects have only a limited (or null) usage for archaeological dating.

Regardless of what scientific techniques are used to obtain the dates, there is in general a great deal of careful thought required in order to link the dates of the objects to the archaeological dating problem under study. A process of logic is needed to link the dates of the objects submitted for (radiocarbon) dating with the corresponding archaeological context or phenomena, see Figure 4.1. This is a crucial factor to be considered in archaeological dating since, without such

linkage, the dates for the objects are of only limited use to archaeologists.

As a further complication, uncertainties can appear at each and every step of the archaeological dating process. There are uncertainties in the radiocarbon dating process, in the chronological relationships between contexts and in the evaluation of the relationships between the dated objects and the archaeological contexts or phenomena under study. It is here that reliable statistical techniques have a crucial rôle in archaeological dating, where it is necessary to measure those uncertainties and combine the different available sources of evidence in a coherent way. As a result of this, we see that the problem of dating in archaeology cannot be viewed simply as the application of a given tool or technique. On the contrary, it must be viewed as a combination of several methods and techniques and as an interdisciplinary and cooperative field of work for archaeologists, physicists, chemists, historians and statisticians (among others).

In Chapter 2 we have examined some archaeological dating problems that illustrate the problems outlined above. In Section 2.4.5 we have reviewed the problem of fixing a floating chronology using radiocarbon dating. Weninger (1986) comments that the principal difficulty is that,

"Cultural phenomena known by Comparative Stratigraphy to be sequential appear to overlap on the ^{14}C scale."

A chronological sequence known *a priori* by stratigraphy, appears not to be in accordance with the radiocarbon determinations available. Radiocarbon determinations, viewed in an isolated way, have little meaning within the corresponding archaeological dating problem at hand. Only when, through a process of calibration, the relationships between the radiocarbon determinations (and thus between the objects radiocarbon dated) and the stratigraphic sequence is

understood and properly modelled, can we use the determinations to date the chronology. (We return to consider this problem in Section 5.3.)

As another example we mentioned the archaeological dating problem examined by Helskog and Schweder (1989) (see Section 2.4.7). In this case the problem was to 'estimate the number of contemporaneous houses at any given time t ' from a site in Norway. A detailed model was used to include in the statistical analysis the relationships between the objects radiocarbon dated and the occupation and longevity of the houses. Without such^a detailed model, the number of contemporaneous houses cannot be inferred from the radiocarbon determinations.

In order to avoid the semantic difficulties of using terms like archaeological 'contexts', 'phases', 'events', we will use the general term 'archaeological dating *phenomena*' or simply '*phenomena*'. The actual definition of such *phenomena* will depend on the particular problem under study. The framework to be constructed in the next Section is only of use when such archaeological dating *phenomena* are properly defined and their relationship with the dated objects is well understood. It is intended that the framework will provide,

- (i) a modelling approach to the calibration of radiocarbon determinations,
- (ii) an approach to the statistical modelling of the archaeological *phenomena* under study and to the statistical modelling of the relationships between the *phenomena* and the objects dated,
- (iii) an interpretation of the usage of probability and statistical models within the context of archaeological dating

and

- (iv) a series of techniques for the implementation of the Bayesian paradigm that provide us with a methodology that consistently combines (i) and (ii), to obtain results coherent in the light of (iii).

4.3 The statistical framework

As we stated above, we want to develop a general framework within which we can tackle a wide variety of statistical problems arising in the interpretation of radiocarbon determinations. We will proceed as follows. In the next Section we present a modelling approach to develop the models needed for the statistical analysis. We do this in the light of the archaeological dating problem discussed above. We then identify the posterior distributions to be calculated and give an interpretation of probability within this context. Later we present a series of techniques to calculate the posterior distributions needed and, in Section 4.3.5, present an example to illustrate how this framework may be applied. Further illustrative examples will be presented in Chapter 5.

4.3.1 Modelling approach

Now we try to clarify the above ideas and express them in terms of probabilistic modelling. Suppose that we have a set of radiocarbon determinations $y_1 \pm \sigma_1, y_2 \pm \sigma_2, \dots, y_m \pm \sigma_m$ and suppose that, given their associated calendar years $\theta_1, \theta_2, \dots, \theta_m$ (ie. θ_j represents the calendar year in which the organic material contained in sample j died), the determinations are independent. Following the calibration process presented in Chapter 3 we have that

$$y_j \mid \theta_j \sim N(\mu(\theta_j), \omega_j^2(\theta_j)), \quad (4.1)$$

where $\omega_j^2(\theta_j) = \sigma_j^2 + \sigma^2(\theta_j)$ (see Equation 3.14) and, since given the θ_j 's the

radiocarbon determinations are independent, we have

$$f(y \mid \theta) = \prod_{j=1}^m f(y_j \mid \theta_j)$$

where $y = (y_1, y_2, \dots, y_m)$ and $\theta = (\theta_1, \theta_2, \dots, \theta_m)$.

The approach we propose for our modelling is to suppose that the prior distribution of each θ_j is in a parametric form given by a vector of parameters $\psi = (\psi_1, \psi_2, \dots, \psi_n)$. The ψ_i 's will then be used to model the characteristics of the archaeological *phenomena* under study. The ψ_i 's can represent a wide variety of aspects depending on the *phenomena* to hand. They could represent, for example, boundaries of some archaeological phases, dates in the past, the elapsed time between particular events, etc. We will then represent the known relationships between the θ_j 's and the *phenomena* by stating the conditional prior distribution

$$f(\theta \mid \psi).$$

With the above modelling approach we attempt to overcome one of the most common sources of confusion in the interpretation of radiocarbon determinations. In Chapters 1 and 2 we have seen that a process of calibration must be taken into account before the radiocarbon determinations can be of use for archaeological dating. Without calibration the relationship between the radiocarbon determinations and the *phenomena* appears to be unclear, perhaps ambiguous or even contradictory. On the other hand, there is archaeological information about the relationship between the samples radiocarbon dated and the archaeological *phenomena*. This information can normally be used to elicit the relationship between the associated calendar years θ_j 's (of the samples) and the *phenomena*. Thus, only through a process of calibration given by the model $f(y \mid \theta)$, is the relationship between the radiocarbon determinations and the archaeological *phenomena* established, when the distribution $f(\theta \mid \psi)$ is stated.

Apart from the above modelling we may have further prior information about the *phenomena* such as a maximum time-span for a phase or an absolute position in the calendar scale for an event, which may be translated into a prior distribution for ψ . Denote this prior distribution by $f(\psi)$.

Given the above modelling approach, the problem of archaeological dating is that of finding the posterior distribution of ψ namely $f(\psi | y)$ and, if of any relevance, the posterior distribution of θ namely $f(\theta | y)$. These two distributions are obtained, using Bayes' theorem, from the joint posterior distribution of the two parameters. That is

$$f(\theta, \psi | y) \propto f(y | \theta, \psi)f(\theta, \psi)$$

(see Section 4.3.4 for a more detailed discussion).

Since the archaeological *phenomena* under study are represented by ψ , the posterior distribution $f(\psi | y)$ will then represent the current knowledge about the position in the calendar scale of the *phenomena*. This is a result of

- (i) the basic modelling assumptions (calibration) involved in radiocarbon dating represented by the model $f(y | \theta)$,
- (ii) the relationship between the associated calendar years (and thus, of the radiocarbon determinations) and the archaeological *phenomena* under study, represented by the distribution $f(\theta | \psi)$,
- (iii) the relevant prior information available about the archaeological *phenomena* represented by the prior distribution $f(\psi)$

and

(iv) the consistent combination of (i), (ii) and (iii) given by the Bayesian framework.

To clarify and simplify the modelling process presented here, we find it useful to represent dependencies (hierarchies) of parameters using a graphical form. In Figure 4.2 we find the 'hierarchy diagram' of our modelling approach explained above.

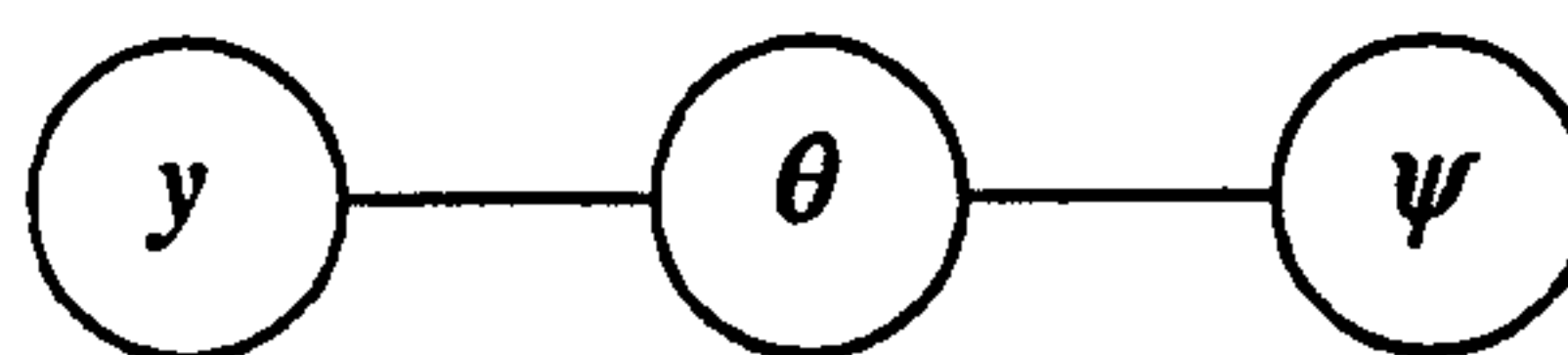


Figure 4.2
Hierarchy diagram of the modelling approach
for our framework.

The diagram in Figure 4.2 tells us that the full conditional distribution of a parameter (or vector of parameters) depends only on the parameters having nodes directly linked to that parameter's node. That is, the full conditional distribution of ψ , $f(\psi | y, \theta)$, is actually equal to $f(\psi | \theta)$ since the corresponding node for ψ is not linked to the corresponding node for y .

Our modelling approach to the archaeological dating problem is illustrated in the above diagram. The relationship the radiocarbon determinations maintain with the archaeological *phenomena* under study is stated only through the associated calendar years θ_j 's. Thus, only the distributions $f(y | \theta)$ and $f(\theta | \psi)$ are part of our model. The unclear *a priori* relationship between y and ψ is of no concern to us. *A posteriori* this relationship is clarified, and archaeological dating is achieved, when Bayes' theorem is applied to calculate $f(\psi | y)$. This has a direct interpretation within the general archaeological dating problem, in which a process of logic is needed to link the objects radiocarbon dated with the *phenomena*

before archaeological dating can be achieved (see Figure 4.1). Such logical linkage will be introduced in our modelling through the specification of $f(\theta | \psi)$.

Besides telling us the corresponding dependencies, hierarchy diagrams give us a quick and clear means of viewing our model. At this stage the advantages of using such 'hierarchy diagrams' do not seem obvious since, perhaps, the modelling process and dependencies were already clear. Nevertheless, as our models increase in complexity and a more detailed description of the relationships between the parameters is needed, we believe that hierarchy diagrams will prove their usefulness and benefit our understanding.

4.3.2 Interpretation of probability within the dating problem

From the above modelling approach we see that θ_j is taken to represent 'the calendar date at which the organic material contained in sample j died'. The particular definition of ψ depends on the archaeological dating *phenomena* under study, but in general the ψ_i 's are taken to represent well defined dates or lengths of time (in the calendar scale) that are part of the *phenomena*. After having defined both θ and ψ we continue to build our model by defining the prior distributions $f(\theta | \psi)$ and $f(\psi)$. Using Bayes' theorem we then calculate $f(\psi | y)$ and $f(\theta | y)$ and, based on those probability distributions, we derive our conclusions. But, what is the interpretation for the (prior or posterior) distributions of θ and ψ ? Or, in general,

What does the probability distribution for a calendar date in the past mean?

For certain definitions of probability the answer to the above question may seem difficult. It might even seem that the idea of having a probability

distribution for a calendar date in the past is absurd. Nothing like that happens if we adopt a subjective definition of probability, as is the case here since we are working within the Bayesian framework. The interpretation of probability within the dating problem is quite straightforward using a subjectivist approach, in which probability is taken to represent degrees of knowledge or belief about uncertainties (see DeGroot 1970, Berger 1985, de Finetti 1974). Our interpretation is as follows.

Even though we consider each θ_j to be a random variable, we do not consider the calendar year θ_j to be intrinsically random in any sense, but rather we consider the probability distribution of θ_j to represent the knowledge we have about this calendar date in the light of present knowledge. Then the radiocarbon determination y_j improves our (prior) knowledge about θ_j by means of computing the distribution of θ_j given y_j via Bayes' theorem. A similar interpretation should apply to ψ depending on the specific archaeological dating *phenomena* under study. Therefore, within this framework, the probability distribution for a parameter representing a calendar date in the past is a measure of the knowledge we have about such a date.

In the trivial case, there may be dates known with complete precision and thus known with probability 1. For example, I know with probability 1 that World War II ended in 1945 AD. In more realistic cases of archaeological dating, we may have a complete probability distribution representing a calendar year in the past. For example, *a priori* we may know that the organic material contained in a given object died at any time θ before the arrival of Columbus in the American continent and after the 11th Century AD. Since we have no further information about the position of θ within this period it seems then reasonable to represent such prior knowledge about θ with an uniform distribution from 450 (1500 AD) to 950 BP (1000 AD). Suppose that we then submit the object for

radiocarbon dating and we receive the determination 700 ± 60 bp. After that, using the calibration procedure explained in Section 1.2, we calculate the corresponding posterior distribution for θ (see Figure 4.3).

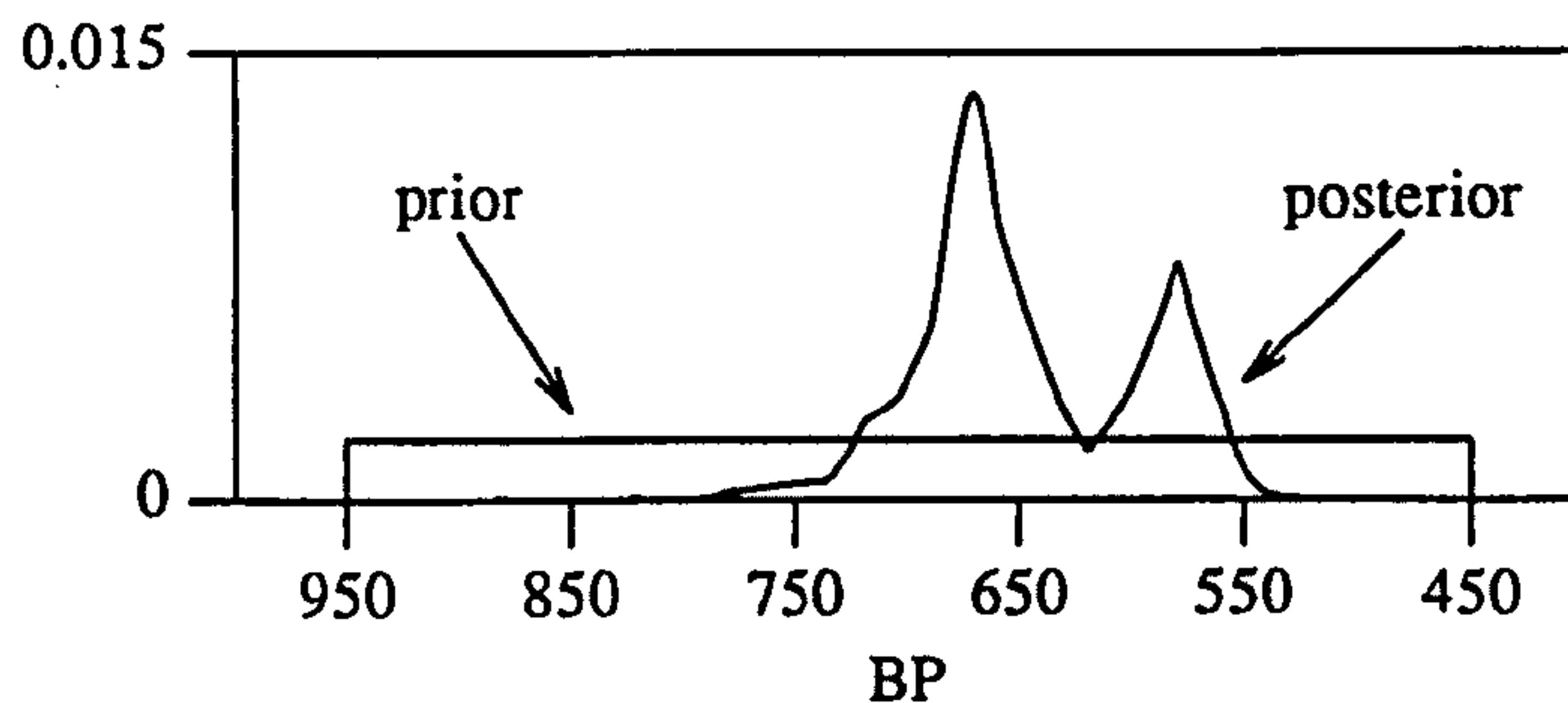


Figure 4.3
Prior distribution for θ , uniform in 450, 950 BP
and the corresponding posterior distribution
given determination 700 ± 60 .

Our knowledge about θ has been modified and improved in the sense that we now, *a posteriori*, know with probability near to 1 that θ should belong to the period of time 770 to 540 BP. Furthermore, we know that the shortest period in the calendar scale with probability 0.95 where θ should be is 740 to 550 BP (that is, 1210 to 1400 AD). (In this artificial example we used the calendar year 670 BP, ie. 1280 AD, to 'create' the 'determination' 700 ± 60 , noticing that $\mu(670) = 700$.)

4.3.3 Consistency

Several authors have discussed the 'consistency' of the Bayesian framework (Jeffreys 1961, de Finetti 1974, DeGroot 1970, Berger 1985). Consistency in this context means non self-contradictory. The Bayesian framework works within the axioms of probability, which are known to be consistent. In a given statistical problem, provided we can represent our knowledge in terms of probability

distributions, we will be consistent.

In the last Section we discussed that, within the archaeological dating problem, there are uncertainties about the position in time of some specific archaeological *phenomena*. DeGroot (1970) proved (chapter 6), under mild conditions, that if our ‘uncertainties’ follow some basic and simple rules of ‘rational behaviour’, then there exists a probability distribution to describe those ‘uncertainties’. In our context this implies that, if our uncertain knowledge about the position in time of the archaeological *phenomena* follows some basic rules of rationale, we must be able to represent such knowledge with a probability distribution. Such^a representation will then be consistent since the axioms of probability are so.

If we then observe the value of some of the ‘unknowns’ (evidence), like the value of a particular radiocarbon determination, Bayes’ theorem tells us how to evaluate the effect this has on the rest of the ‘unknowns’. Thus for Bayesian statistics, learning from data in a consistent way comes solely as a result of working within the axioms of probability, since Bayes’ theorem is a consequence of those axioms. Thus, the statistical framework here presented is consistent simply because the Bayesian framework is so.

4.3.4 Posterior distributions and MCMC methods

In this Section we will concentrate on the implementation of the Bayesian paradigm within our statistical framework. That is, calculate the posterior distributions of interest. From the above discussion on modelling we see that we have a hierarchical model using the vectors of parameters ψ and θ and the data (the radiocarbon determinations) y . The density $f(y \mid \theta, \psi)$ is equal to $f(y \mid \theta) = L(y \mid \theta)$, the likelihood, and from Equation 4.1 we see that

$$L(y | \theta) = \prod_{j=1}^m L(y_j | \theta_j) \propto \prod_{j=1}^m \frac{1}{\omega_j(\theta_j)} \exp \left\{ -\frac{1}{2} \left(\frac{(y_j - \mu(\theta_j))^2}{\omega_j^2(\theta_j)} \right) \right\},$$

where $\omega_j^2(\theta_j) = \sigma_j^2 + \sigma^2(\theta_j)$. Also we have the prior densities $f(\theta | \psi)$ and $f(\psi)$. As stated before we need $f(\psi | y)$ and, if relevant, $f(\theta | y)$ (these are the posterior densities of ψ and θ) which will be obtained by applying Bayes' theorem. The approach we take is to calculate $f(\theta, \psi | y)$ and from that integrate out ψ or θ to obtain $f(\theta | y)$ or $f(\psi | y)$, respectively. In more detail, Bayes' theorem states that

$$f(\theta, \psi | y) = \frac{f(y | \theta, \psi) f(\theta, \psi)}{\iint f(y | \theta, \psi) f(\theta, \psi) d\theta d\psi}.$$

But $f(y | \theta, \psi) = L(y | \theta)$ and $f(\theta, \psi) = f(\theta | \psi) f(\psi)$, and therefore

$$f(\theta, \psi | y) = \frac{L(y | \theta) f(\theta | \psi) f(\psi)}{\iint L(y | \theta) f(\theta | \psi) f(\psi) d\theta d\psi}.$$

From this we have

$$f(\psi | y) = \int \frac{L(y | \theta) f(\theta | \psi) f(\psi)}{\iint L(y | \theta) f(\theta | \psi) f(\psi) d\theta d\psi} d\theta$$

and

$$f(\theta | y) = \int \frac{L(y | \theta) f(\theta | \psi) f(\psi)}{\iint L(y | \theta) f(\theta | \psi) f(\psi) d\theta d\psi} d\psi.$$

$f(\psi | y)$ and $f(\theta | y)$ are difficult to interpret since they are multidimensional distributions. It is commonly more useful to obtain the marginal distributions of the θ_j 's and the ψ_i 's, which are

$$f(\psi_i | y) = \int \cdots \iint f(\psi | y) d\psi_1 \cdots d\psi_{i-1} d\psi_{i+1} \cdots d\psi_n,$$

with a similar expression for $f(\theta_j | y)$.

The above integrals are often difficult to evaluate. In fact, analytical methods are discarded since $L(y | \theta)$ (as seen above) includes both $\mu(\theta)$ and $\sigma^2(\theta)$

which both have a numerical definition (piece-wise linear and piece-wise quadratic, see Equations 1.3 and 3.14). Hence numerical methods need to be used. However, apart from the simplest models, standard numerical methods for integration will be difficult or impossible to implement due to the large dimensionality of the parameters θ and ψ .

Fortunately for us, methods for approximating posterior distributions like the ones above, with complicated forms and high dimensionalities, have recently been developed allowing us to obtain $f(\psi_i | y)$ and $f(\theta_j | y)$ (or nearly any other posterior distribution or posterior moments of interest). These are sampling-resampling Monte Carlo methods commonly known as 'Markov chain Monte Carlo' (MCMC) methods (see Geman and Geman 1984, Gelfand and Smith 1990). Since 1990 there has been important progress in the development of MCMC methods and these have been applied in many areas, solving problems that otherwise would have been impossible to tackle (see Carlin *et al.* 1992, Stephens 1994, Gilks *et al.* 1993). A review of the subject can be found in Smith and Roberts (1993). Here we present a basic explanation of the 'Gibbs sampler', which is a special case of MCMC methods.

Suppose we have an n -dimensional parameter u , whose posterior distribution is denoted by $f(u) = f(u_1, u_2, \dots, u_n)$. We denote the full posterior conditional distributions by

$$f(u_1 | u_2, u_3, \dots, u_n),$$

$$f(u_2 | u_1, u_3, \dots, u_n),$$

...

$$f(u_n | u_1, u_2, \dots, u_{n-1}).$$

The Gibbs sampler involves sampling from the full conditional distributions for the parameters in our model and generating a 'sampling-resampling' scheme to approximate the posterior distribution under investigation $f(u)$ or any of its marginals or moments. The Gibbs sampling scheme consists in choosing initial values for u_2, u_3, \dots, u_n , (say $u_2^{(0)}, u_3^{(0)}, \dots, u_n^{(0)}$). Then $u_1^{(1)}$ is generated from the conditional distribution

$$f(u_1 | u_2^{(0)}, u_3^{(0)}, \dots, u_n^{(0)}).$$

Next $u_2^{(1)}$ is generated from the conditional distribution

$$f(u_2 | u_1^{(1)}, u_3^{(0)}, \dots, u_n^{(0)}).$$

The process continues up to generating $u_n^{(1)}$ from the conditional distribution

$$f(u_n | u_1^{(1)}, u_2^{(1)}, \dots, u_{n-1}^{(1)})$$

and having $u^{(1)} = (u_1^{(1)}, u_2^{(1)}, \dots, u_n^{(1)})$. This process of passing from $u^{(k)}$ to $u^{(k+1)}$ forms a Markov process and Geman and Geman (1984) have shown that, under fairly general conditions

$$u^{(k)} \xrightarrow{d} u \sim f(u) \text{ as } k \rightarrow \infty.$$

It can also be proved that, for any measurable function g

$$\frac{1}{s} \sum_{k=1}^s g(u^{(k)}) \rightarrow E[g(u)] \text{ as } s \rightarrow \infty.$$

Given the above Markov process for large k , $u^{(k)}$ can be seen as a realisation of u . From this we can then use the first of the above results to generate an independent 'sample' for u . Clearly subsequent $u^{(k)}$'s will be correlated. However, this can be avoided by either of two different methods. The first one is, using different starting values $u^{(0)}$ run the chain s times and thus obtain $u_1^{(k)}, u_2^{(k)}, \dots, u_s^{(k)}$ which can be seen as an independent sample for u . The second method is to run a single chain but only consider $u^{(k)}$'s at suitable spacings, this after leaving the chain to run for a large number of iterations ('the burn-in period') to allow for convergence. That is, taking $u^{(k)}$'s for $k = k_1 + iq, i = 1, 2, \dots, s$. This then will avoid correlation and thus we will obtain an independent sample for u . Using either method we can then approximate any marginals or posterior moments for u given the sample just obtained. There is still controversy as to which method should be used (multiple runs or a single run). However, the latest thinking suggests that the latter should be preferred. That is, taking $u^{(k)}$'s for $k = k_1 + iq, i = 1, 2, \dots, s$ as an independent sample for u (see Geyer 1992). This is the sampling scheme we use in our examples.

Now using the second result above we can approximate some moments for u (for example the expectation of u using $g(u) = u$). In this case, though, the $u^{(k)}$'s need not be independent and we can use all the elements in a single chain to approximate the desired moments or expectations $E[g(u)]$.

If we are interested in approximating the marginals for u_i denoted by $f(u_i)$, we may take $u_i^{(k_1+iq)}$ as a sample for u_i . The density of u_i is then approximated with, for example, a histogram of such a sample. However, Gelfand and Smith (1990) argue that a procedure that might be more efficient is to approximate $f(u_i)$ pointwise by averaging the full conditionals of u_i for each $u^{(k_1+iq)}$. That is

$$f(u_i) \approx \frac{1}{s} \sum_{i=1}^s f(u_i \mid u_1^{(k)}, \dots, u_{i-1}^{(k)}, u_{i+1}^{(k)}, \dots, u_n^{(k)})$$

where $k = k_1 + iq$.

A crucial issue in the MCMC methodology is how to select the values of k_1 , q and s , that is the burn-in period, the spacings and the sample size, in order to obtain convergence and proper approximations for the marginals or posterior moments required. In the last few years much research has appeared studying this issue and indeed the values given for k_1 , q and s will depend on the specific problem at hand. Smith and Roberts (1993) list some cases of MCMC implementation for which some convergence properties can be predicted. There are also some methods to monitor convergence while sampling, see Gelman and Rubin (1992), Raftery and Lewis (1992) and Roberts (1992). In general, however, these methods are difficult to apply and the convergence analysis of MCMC methods is commonly carried out by checking for consistency in the results. Smith and Roberts (1993) recommend that we run the whole process with different k_1 , q and s values and different starting values $u^{(0)}$, and verify that our results are similar, thus giving a high degree of confidence that convergence has been achieved.

Before presenting an explanation of how to apply MCMC methods within the framework presented here, we need to introduce some notation. If we have any vector u then, unless clearly stated, u_i is the i^{th} component of such vector and u_{-i} is the vector $(u_1, u_2, \dots, u_{i-1}, u_{i+1}, \dots, u_n)$. Also we will denote the indicator function by $I(x)_A$, which is defined as

$$I(x)_A = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

Using Bayes' theorem, it is easy to see that the full posterior conditional distributions for our framework are

$$f(\theta_j \mid y, \theta_{-j}, \psi) \propto L(y \mid \theta) f(\theta_j \mid \theta_{-j}, \psi), \quad (4.2)$$

for $j = 1, 2, \dots, m$, where $f(\theta_j \mid \theta_{-j}, \psi) \propto f(\theta \mid \psi)$ viewed as a function of θ_j .

Similarly we see that

$$f(\psi \mid y, \theta) \propto L(y \mid \theta)f(\psi \mid \theta).$$

Since we are working with a hierarchical model, $L(y \mid \theta)$ is independent of ψ and then

$$f(\psi \mid y, \theta) = f(\psi \mid \theta) \propto f(\theta \mid \psi)f(\psi)$$

or

$$f(\psi_i \mid y, \psi_{-i}, \theta) \propto f(\theta \mid \psi)f(\psi_i \mid \psi_{-i}) \quad (4.3)$$

for $i=1,2,\dots,n$, where $f(\psi_i \mid \psi_{-i}) \propto f(\psi)$ viewed as a function of ψ_i . Then using these full conditionals for the θ_j 's and the ψ_i 's, we can use a MCMC method (like the Gibbs sampler) to approximate $f(\psi \mid y)$ or any marginal $f(\theta_j \mid y)$ or $f(\psi_i \mid y)$ of interest.

In Equations 4.2 and 4.3 we find the distributions $f(\theta_j \mid \theta_{-j}, \psi)$ and $f(\psi_i \mid \psi_{-i})$. These distributions arise from $f(\theta \mid \psi)$ and $f(\psi)$ respectively. Since in most problems we are going to use MCMC methods it would be more useful to elicit our prior information about θ and ψ in terms of these conditional prior distributions. These distributions are normally easier to state since they are univariate, unlike $f(\theta \mid \psi)$ and $f(\psi)$.

4.3.5 An example

To clarify ideas we give an example of how the framework explained is applied to a specific dating problem. Here we concentrate on the modelling and the techniques needed to calculate the posterior distributions of interest. Using this same model, we later analyse a specific set of radiocarbon determinations.

Suppose that we have a set of radiocarbon determinations $y_1 \pm \sigma_1, y_2 \pm \sigma_2, \dots, y_m \pm \sigma_m$ associated with the (unknown) calendar years $\theta_1, \theta_2, \dots, \theta_m$. Suppose further that we are working with a single archaeological *phenomenon*. This can be, for example, the occupation of a particular archaeological site or the existence of a given culture, with the consequence that the calendar years $\theta_1, \theta_2, \dots, \theta_m$ belong to the time-span of the archaeological *phenomenon*. This is a common problem in archaeology in which there is usually little or no prior information about the internal relationship between the θ_j 's such as, for example, chronological orderings or further internal subgroupings. In such circumstances archaeologists wish to summarise the radiocarbon determinations and make inferences about the *phenomenon*'s time-span and/or its duration, ideally trying to combine the evidence of the radiocarbon determinations with the archaeological information they have about the *phenomenon*. We will call this the 'summarisation problem'.

According to our framework, we need to define ψ and then state the distributions $f(y \mid \theta)$, $f(\theta_j \mid \theta_{-j}, \psi)$ and $f(\psi_i \mid \psi_{-i})$ according to the relationship between the objects radiocarbon dated and the *phenomenon* under study. In this respect, the crucial point is that we know that the objects radiocarbon dated died within the time-span of the *phenomenon* under study. We define $\psi = (\alpha, \beta)$ (here $\psi_1 = \alpha$ and $\psi_2 = \beta$) and then β will represent the beginning and α the end, on the calendar scale, of the *phenomenon* under study (years BP). Now $f(y \mid \theta)$ is given by Equation 4.1 and, since we supposed that we have only vague prior information about internal relationships between the θ_j 's, we have $f(\theta_j \mid \theta_{-j}, \psi) = f(\theta_j \mid \psi)$. One possible definition for $f(\theta_j \mid \psi) = f(\theta_j \mid \alpha, \beta)$ that may be suitable for several specific cases of the summarisation problem is

$$\theta_j \mid \alpha, \beta \sim U(\alpha, \beta); j = 1, 2, \dots, m.$$

That is, the θ_j 's are uniformly distributed over a period of time beginning at β

and finishing at α (years BP). This type of distribution could be slightly generalised by only stating that the distribution of $\theta_j \mid \alpha, \beta$ has support on (α, β) and not restricting it to be uniform. However, this will require more archaeological information to be available. In the light of no further information along these lines we proceed then with the uniform case.

An example of the prior distributions for α and β is

$$\alpha \sim U(a_1, b_1) \text{ and } \beta \sim U(a_2, b_2) \quad (4.4)$$

for some positive $a_1 < b_1 < a_2 < b_2$. That is, uniform distributions with some (wide) margins that do not overlap. Since the distributions do not overlap we can suppose that α and β are independent and thus, $f(\alpha \mid \beta) = f(\alpha)$ and $f(\beta \mid \alpha) = f(\beta)$. However, in some cases it might be necessary to state a more informative prior for α and β depending on the prior information available, possibly with bell-shaped distributions with support on (α, β) .

We now create the hierarchy diagram for the summarisation problem to clarify the whole modelling process. We do it in two stages. The general hierarchy diagram is given in Figure 4.4.

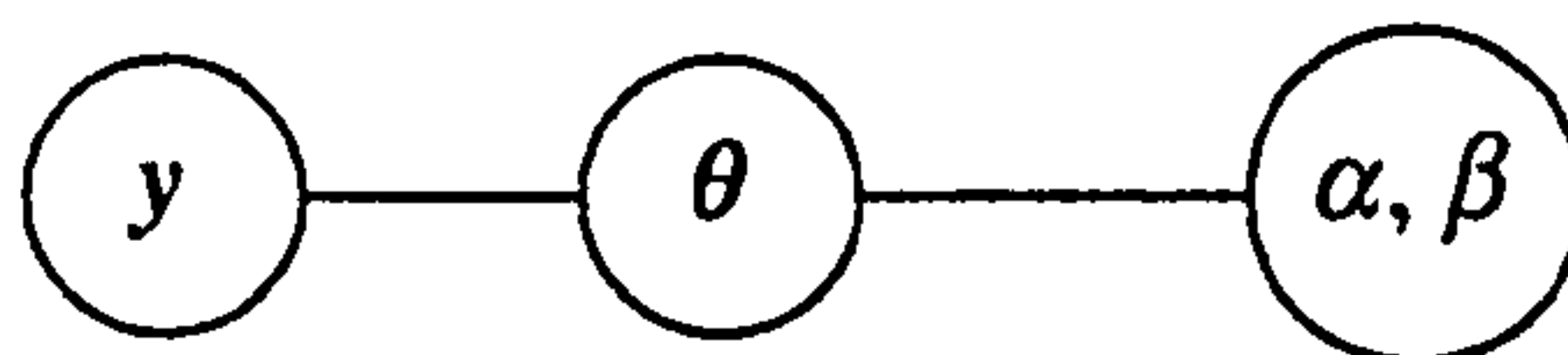


Figure 4.4
General hierarchy diagram for the summarisation problem.

Since we are using the model from Equation 4.1 and we suppose that the y_j 's are independent, a more detailed (and informative) hierarchy diagram for this model can be seen in Figure 4.5. Note that the nodes corresponding to α and β are joined together since the full conditional distribution of α depends on β and vice versa.

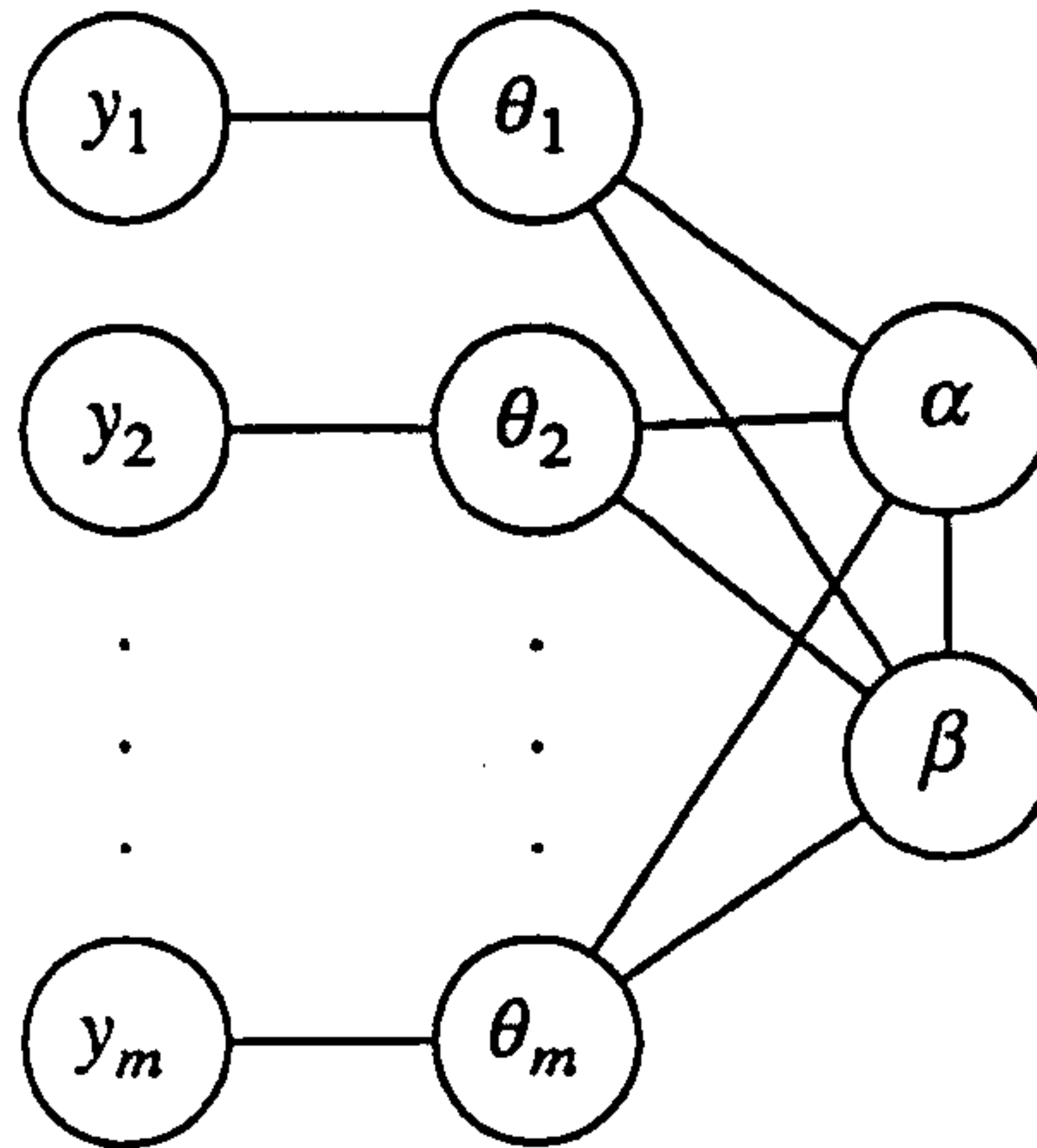


Figure 4.5
Hierarchy diagram for the summarisation problem.

Now, we need the posterior distribution $f(\psi \mid y)$, which in this example is

$$f(\alpha, \beta \mid y).$$

This distribution will then represent the current knowledge about the time-span (β to α) of the archaeological *phenomenon* under study given the radiocarbon determinations y and the prior distributions for α and β .

Using Equations 4.2 and 4.3 we see that the full conditional distributions of the parameters in the summarisation problem are

$$f(\theta_j \mid y, \theta_{-j}, \alpha, \beta) = f(\theta_j \mid y_j, \alpha, \beta) \propto \frac{1}{\omega_j(\theta_j)} \exp \left\{ -\frac{1}{2} \left(\frac{(y_j - \mu(\theta_j))^2}{\omega_j^2(\theta_j)} \right) \right\} I(\theta_j)_{(\alpha, \beta)},$$

$$f(\alpha \mid y, \theta, \beta) = f(\alpha \mid \theta, \beta) \propto (\beta - \alpha)^{-m} I(\alpha)_{(0, \min(\theta))} f(\alpha \mid \beta),$$

$$f(\beta \mid y, \theta, \alpha) = f(\beta \mid \theta, \alpha) \propto (\beta - \alpha)^{-m} I(\beta)_{(\max(\theta), \infty)} f(\beta \mid \alpha)$$

where $\max(\theta)$ and $\min(\theta)$ are the maximum and the minimum of the components

of θ respectively.

If we take $f(\alpha)$ and $f(\beta)$ as the priors in Equation 4.4, we can then suppose $f(\alpha, \beta) = f(\alpha)f(\beta)$ and the full conditionals for α and β will be

$$f(\alpha \mid \theta, \beta) \propto (\beta - \alpha)^{-m} I(\alpha)_{(a_1, b)}$$

and

$$f(\beta \mid \theta, \alpha) \propto (\beta - \alpha)^{-m} I(\beta)_{(a, b_2)},$$

where $b = \min(\min(\theta), b_1)$ and $a = \max(\max(\theta), a_2)$.

Once having the above distributions we can implement a MCMC method, like the Gibbs sampler, to approximate the posterior distributions of α and β and, therefore, find the limits for the time-span of the archaeological *phenomenon* under study. To sample from $\alpha \mid \theta, \beta$ we first obtain $K^{-1} = \int_{a_1}^b (\beta - \alpha)^{-m} d\alpha$ (thus K is the normalisation constant). Second we obtain $F_{\alpha \mid \theta, \beta}(\alpha^*) = \int_{a_1}^{\alpha^*} K (\beta - \alpha)^{-m} d\alpha$. From that we solve the equation $F_{\alpha \mid \theta, \beta}(\alpha^*) = u$ to obtain

$$\alpha^* = \beta - \frac{\beta - b}{\left\{ \left(\frac{\beta - b}{\beta - a_1} \right)^{m-1} (1-u) + u \right\}^{\frac{1}{m-1}}}.$$

Then, simulating a value, u , from the uniform distribution $U(0, 1)$ will give us a simulated value, α^* , for $\alpha \mid \theta, \beta$. Similarly for $\beta \mid \theta, \alpha$ we see that

$$\beta^* = \alpha - \frac{a - \alpha}{\left\{ \left(\frac{a - \alpha}{b_2 - \alpha} \right)^{m-1} (1-u) + u \right\}^{\frac{1}{m-1}}}$$

where $U \sim U(0, 1)$. However, to sample from $f(\theta_j \mid y_j, \alpha, \beta)$ we need a numerical method since such distributions cannot be stated analytically (given the definitions of $\mu(\theta)$ and $\sigma^2(\theta)$). Several methods can be applied, for example,

rejection or inversion methods. In our calculations we choose the method of inversion by numerical solution. That is, to normalise and use numerical integration to find the value θ^* such that $F_{\theta_j | y_j, \alpha, \beta}(\theta^*) = u$, where again $U \sim U(0, 1)$ (see Devroye 1986 for further details).

In Section 5.2 we analyse a specific example of the summarisation problem with radiocarbon determinations related to a pre-Hispanic culture in Peru. We will see that the framework developed here is relatively simple to implement and is able to provide meaningful answers to the archaeological questions posed.

4.4 Replications

In this Section we study the dating problem in which we have a set of replicated radiocarbon determinations. The statistical analysis of replications fits within our framework and represents an important, although technical, problem that needs to be addressed. We have introduced this problem in Section 2.2.2, where we reviewed other approaches. Here we present our own approach.

The problem is as follows. We have a set of radiocarbon determinations $\{y_1 \pm \sigma_1, y_2 \pm \sigma_2, \dots, y_m \pm \sigma_m\}$ all associated with a single calendar year θ . That is, we are assuming that θ is the calendar year in which the organic material contained in the m samples radiocarbon dated died. The main interest in this problem is to use the radiocarbon determinations to increase our knowledge about the position in the calendar scale of θ . In the next Section we will study a replications dating problem in which this is the case. In general, however, the problem of replications is not considered on its own but as part of a more complex dating problem. For this reason it is of prime interest for us to have a correct approach to this basic problem.

In the light of the statistical framework presented in this Chapter, for the case of replications, the parameter ψ will be non-random letting $f(\theta) = f(\theta | \psi)$ be the prior distribution for θ . Given θ we suppose that the determinations are independent and thus $L(y | \theta) = \prod_{j=1}^m f(y_j | \theta)$. In this case our analysis consists of finding

$$f(\theta | y) \propto L(y | \theta)f(\theta).$$

That is, the posterior distribution of θ given the determinations y . We present the corresponding hierarchy diagram for this case in Figure 4.6.

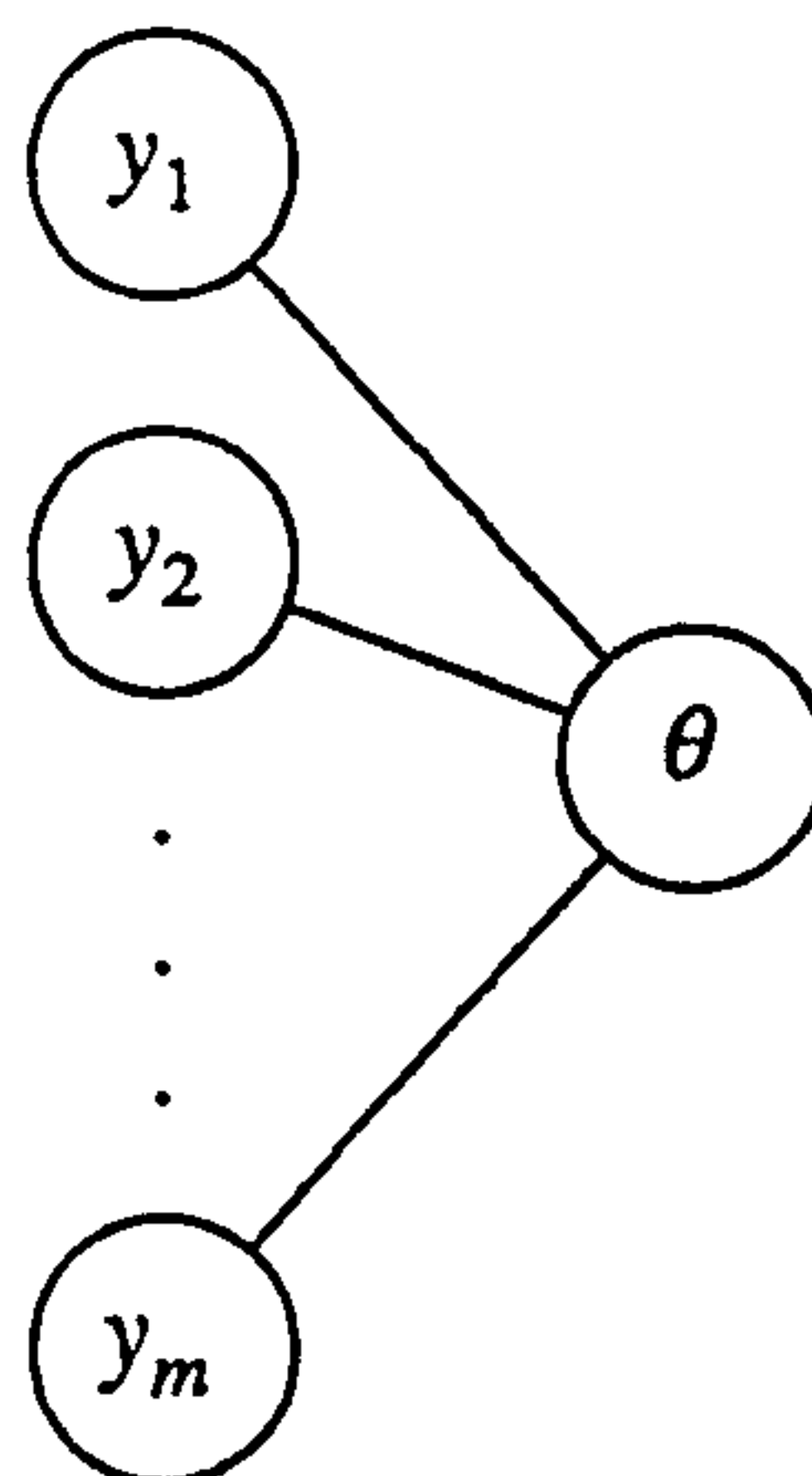


Figure 4.6
Hierarchy diagram for the replications problem.

In the interpretation of replicated radiocarbon determinations it is common practice among the radiocarbon community to use the pooled mean to summarise the determinations and make any subsequent inferences. Using the conventional calibration model explained in Chapter 1, we will prove below that the pooled mean is a sufficient statistic. However, we will see that using the new calibration procedure expounded in Chapter 3 (considering the variance in the calibration curve), the pooled mean is not necessarily a sufficient statistic.

Using the calibration procedure presented in Chapter 1 and using Equation 1.4 we see that,

$$y_j \mid \theta \sim N(\mu(\theta), \sigma_j^2)$$

for $j = 1, 2, \dots, m$ and therefore the likelihood is

$$L(y \mid \theta) = \left\{ \prod_{j=1}^m (2\pi\sigma_j^2)^{-\frac{1}{2}} \right\} \exp \left\{ -\frac{1}{2} \sum_{j=1}^m \frac{(y_j - \mu(\theta))^2}{\sigma_j^2} \right\}.$$

If we complete the square in the expression $\sum_{j=1}^m \frac{(y_j - \mu(\theta))^2}{\sigma_j^2}$ it is easy to see that

$$L(y \mid \theta) = \left\{ \prod_{j=1}^m (2\pi\sigma_j^2)^{-\frac{1}{2}} \right\} \exp \left\{ -\frac{1}{2} C(y) \right\} \exp \left\{ -\frac{1}{2} \left(\frac{(m_y - \mu(\theta))^2}{\sigma^2} \right) \right\} \quad (4.5)$$

where

$$\sigma^{-2} = \sum_{j=1}^m \frac{1}{\sigma_j^2}, \quad m_y = \sigma^2 \sum_{j=1}^m \frac{y_j}{\sigma_j^2}, \quad \text{and} \quad C(y) = \sum_{j=1}^m \frac{(y_j)^2}{\sigma_j^2} - \sigma^2 \left(\sum_{j=1}^m \frac{y_j}{\sigma_j^2} \right)^2.$$

Using the factorisation theorem (Mood *et al.* 1974, p. 307), we deduce that the pooled mean, m_y , is a sufficient statistic for y . Given a prior distribution for θ , $f(\theta)$, we then have

$$f(\theta \mid y) = f(\theta \mid m_y) \propto \exp \left\{ -\frac{1}{2} \left(\frac{(\mu(\theta) - m_y)^2}{\sigma^2} \right) \right\} f(\theta).$$

This means that, using the above model in which we neglect the variance in the calibration curve, we can use the pooled mean as a summary for our replications sample and not lose information. However, when we turn to consider the variance in the calibration curve this is not necessarily the case. The model now is

$$y_j \mid \theta \sim N(\mu(\theta), \omega_j^2(\theta)) \quad (4.6)$$

where $\omega_j^2(\theta) = \sigma_j^2 + \sigma^2(\theta)$ (see Equation 3.14). The likelihood will have a similar expression to the one above but substituting $\omega_j^2(\theta)$ for σ_j^2 . But notice now that

the factorisation of $L(y | \theta)$ found above does not imply the sufficiency of the pooled mean since $C(y)$ will depend on θ (actually the proof is not applicable at all since in this case $m_y = \sigma^2 \sum_{j=1}^m \frac{y_j}{\omega_j^2(\theta)}$ is not the pooled mean). To prove that the pooled mean is not necessarily a sufficient statistic using this model, we consider a counterexample. We take the determinations 1200, 1230 and 1250, all with standard deviation $\sigma = 20$, and calculate the posterior distribution of the associated calendar year θ using this latter model (and a vague prior distribution for θ). That is,

$$f(\theta | y) \propto \prod_{j=1}^m \frac{1}{\omega_j(\theta)} \exp \left\{ -\frac{1}{2} \frac{(y_j - \mu(\theta))^2}{\omega_j^2(\theta)} \right\}$$

where $m = 3$. We then calculate $f(\theta | m_y)$ to compare it with $f(\theta | y)$ (if m_y is sufficient then these two densities should be equal). To do this we need to calculate the distribution $f(m_y | \theta)$. We know that $m_y | \theta$ is a weighted average of normally distributed random variables (the $(y_j | \theta)$'s) and thus $m_y | \theta$ has a normal distribution. Given that in this case we have equal variances, it is easy to see that $m_y | \theta \sim N(\mu(\theta), \sigma^2/m + \sigma^2(\theta)/m)$. Using this we then compute $f(\theta | m_y)$, with a vague prior distribution for θ (the pooled mean is 1226.6 with standard deviation 11.5).

In Figure 4.7 we display histograms of both $f(\theta | y)$ and $f(\theta | m_y)$, and since they are not identical, we conclude that the pooled mean is not necessarily a sufficient statistic. However, only if the variances σ_j^2 's are small will there be a noticeable difference in the distributions. Otherwise we could suppose that $\sigma_j^2 = \sigma_j^2 + \sigma^2(\theta)$ and thus the pooled mean will be sufficient.

The above analysis tells us that, potentially, we might be losing information when using the pooled mean to summarise a set of replicated radiocarbon determinations. However, this will only occur when we have determinations with very low standard deviations. In any case, to avoid a potential loss of important

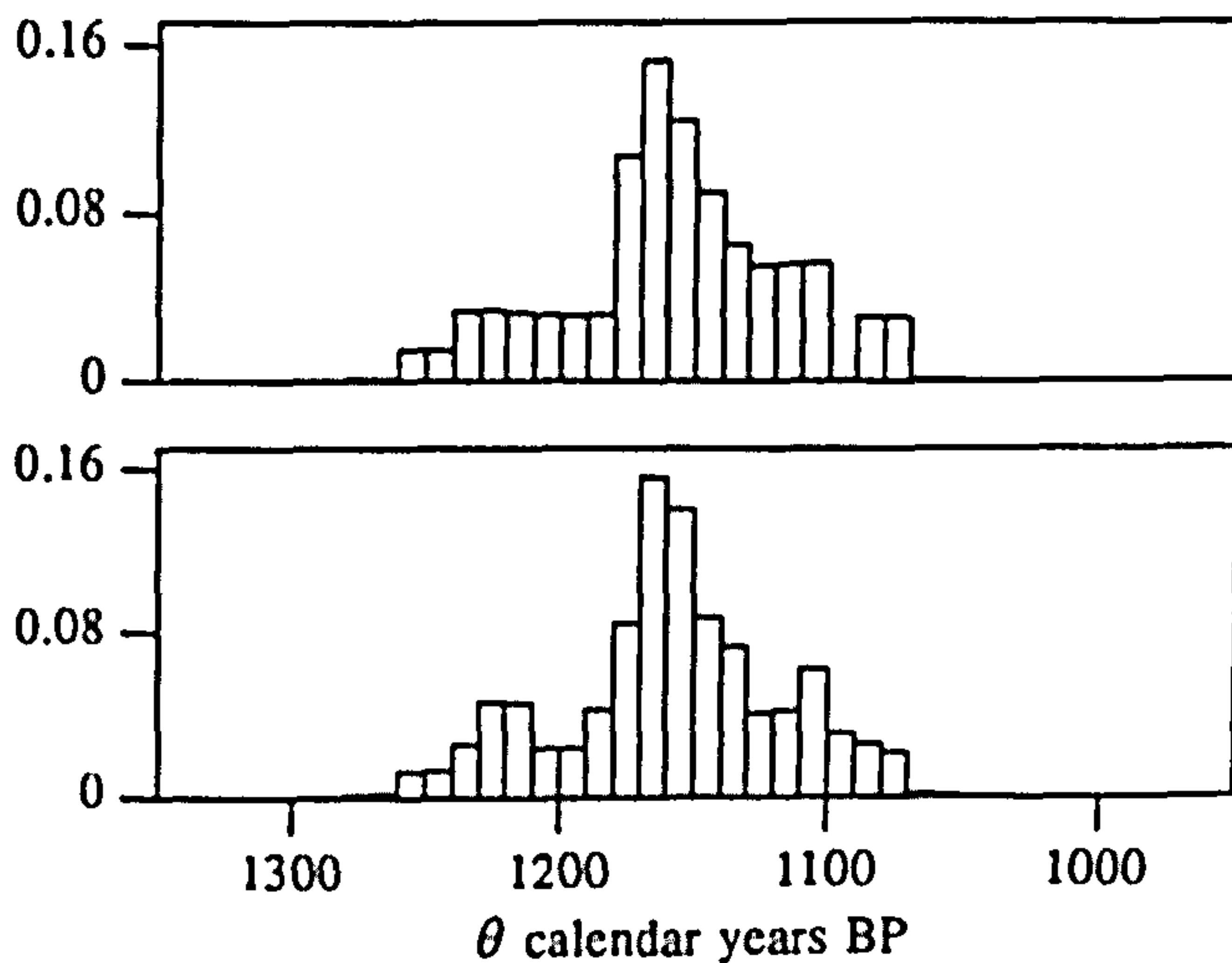


Figure 4.7

Histograms for $\theta | y$ (above) and $\theta | m_y$ (below), considering the errors in the calibration curve, for the determinations 1200, 1230 and 1250 with $\sigma = 20$.

Since the histograms are not identical, m_y (the pooled mean) is not necessarily a sufficient statistic.

information, we would recommend reporting the actual radiocarbon determinations and not only the pooled mean.

Using a technique presented by Lehmann and Scheffé (1950) it is not difficult to prove that, if the standard deviations σ_j 's are equal, then $\left(\sum_{j=1}^m y_j, \sum_{j=1}^m y_j^2\right)$ is a sufficient statistic. This technique consists of analysing the ratio $f(y | \theta)/f(x | \theta)$. Then any statistic $t(\cdot)$ such that $t(y) = t(x)$ if and only if $f(y | \theta)/f(x | \theta)$ is not dependent upon θ , is sufficient (and, actually, minimal). If we take $y_j \sim N(\mu(\theta), \omega_j^2(\theta))$ and $x_j \sim N(\mu(\theta), \omega_j^2(\theta))$ for $j = 1, 2, \dots, m$ then

$$\frac{f(y | \theta)}{f(x | \theta)} = \exp \left\{ -\frac{1}{2} \left(\sum_{j=1}^m \frac{(y_j^2 - x_j^2) - 2\mu(\theta)(x_j - y_j)}{\omega_j^2(\theta)} \right) \right\}. \quad (4.7)$$

Now, if the standard deviations σ_j 's are equal, and thus $\omega_j^2(\theta) = \omega^2(\theta)$, we have

$$\frac{f(y | \theta)}{f(x | \theta)} = \exp \left\{ -\frac{1}{2\omega^2(\theta)} \left(\left(\sum_{j=1}^m y_j^2 - \sum_{j=1}^m x_j^2 \right) - 2\mu(\theta) \left(\sum_{j=1}^m y_j - \sum_{j=1}^m x_j \right) \right) \right\}.$$

Thus $\sum_{j=1}^m y_j = \sum_{j=1}^m x_j$ and $\sum_{j=1}^m y_j^2 = \sum_{j=1}^m x_j^2$ if and only if $f(y | \theta)/f(x | \theta)$ is not dependent upon θ and therefore

$$\left(\sum_{j=1}^m y_j, \sum_{j=1}^m y_j^2 \right)$$

is a sufficient statistic. From Equation 4.7 it seems, however, that if the standard errors σ_j 's are not equal the only sufficient statistic available would be the sample itself. The latter is commonly the case for sets of replicated samples and consequently the above result will be of only limited use.

4.4.1 Example: The Shroud of Turin

Since the development of AMS radiocarbon dating, which only requires very small amounts of organic material (milligrammes), it has been possible to test the authenticity or otherwise of the 'Shroud of Turin' (widely believed to have been used to wrap the body of Christ). More technical details may be found in Damon *et al.* (1989b).

Our interest in this problem is that the samples all come from a single object (the Shroud) and therefore we can state that all twelve radiocarbon analyses performed on it should relate to the same calendar year. A strip (approximately 10×70mm) was cut from the Shroud and divided into three samples which were sent, along with control samples, to the Arizona, Oxford and Zurich AMS laboratories to be radiocarbon dated. The laboratories made four, three and five determinations respectively (see Table 4.1).

The Shroud of Turin was first displayed in the 1350's AD (ie. about 600 BP, Damon *et al.* 1989b). Therefore the organic matter in the Shroud should have died before that date and thus we state the lower boundary of 600 BP for the support of the prior distribution of θ . We decided to state a uniform prior

| Laboratory & sample id. | Sample (rep.) | Det. |
|----------------------------|------------------|--------------|
| Arizona AA-3367 | A1.1 | 591 ± 30 |
| | A1.2 | 690 ± 35 |
| | A1.3 | 606 ± 41 |
| | A1.4 | 701 ± 33 |
| Oxford 2575 | O1.1 | 795 ± 65 |
| | O1.2 | 730 ± 45 |
| | O1.3 | 745 ± 55 |
| Zurich ETH-3883 | Z1.1 | 733 ± 61 |
| | Z1.2 | 722 ± 56 |
| | Z1.3 | 635 ± 57 |
| | Z1.4 | 639 ± 45 |
| | Z1.5 | 679 ± 51 |

Table 4.1

Radiocarbon determinations for the 'Shroud of Turin', from Damon *et. al.* (1989), all are AMS datings using different pretreatments.

distribution for θ , say $\theta \sim U(600, b)$ for some suitable large b (uninformative).

We used all twelve radiocarbon determinations and, employing the procedure explained above, including the variance in the calibration curve, calculated the posterior distribution of θ which is given in Figure 4.8. For comparison, in Figure 4.9 we present the corresponding posterior distribution for θ found when we do not consider the variance in the calibration curve. As explained above, in such a case we can use the pooled mean as a sufficient statistic for the data (the pooled mean in this case is 679 ± 12). Note that, given the magnitude of the standard deviations for the determinations (more than 30 years), the variance in the calibration curve could be neglected and thus there is nearly no difference between the two methods.

Note from Figure 4.8 that the whole of the posterior distribution for θ (100% HPD region) lies between 1270 and 1310 AD (680-640 BP). This means that our

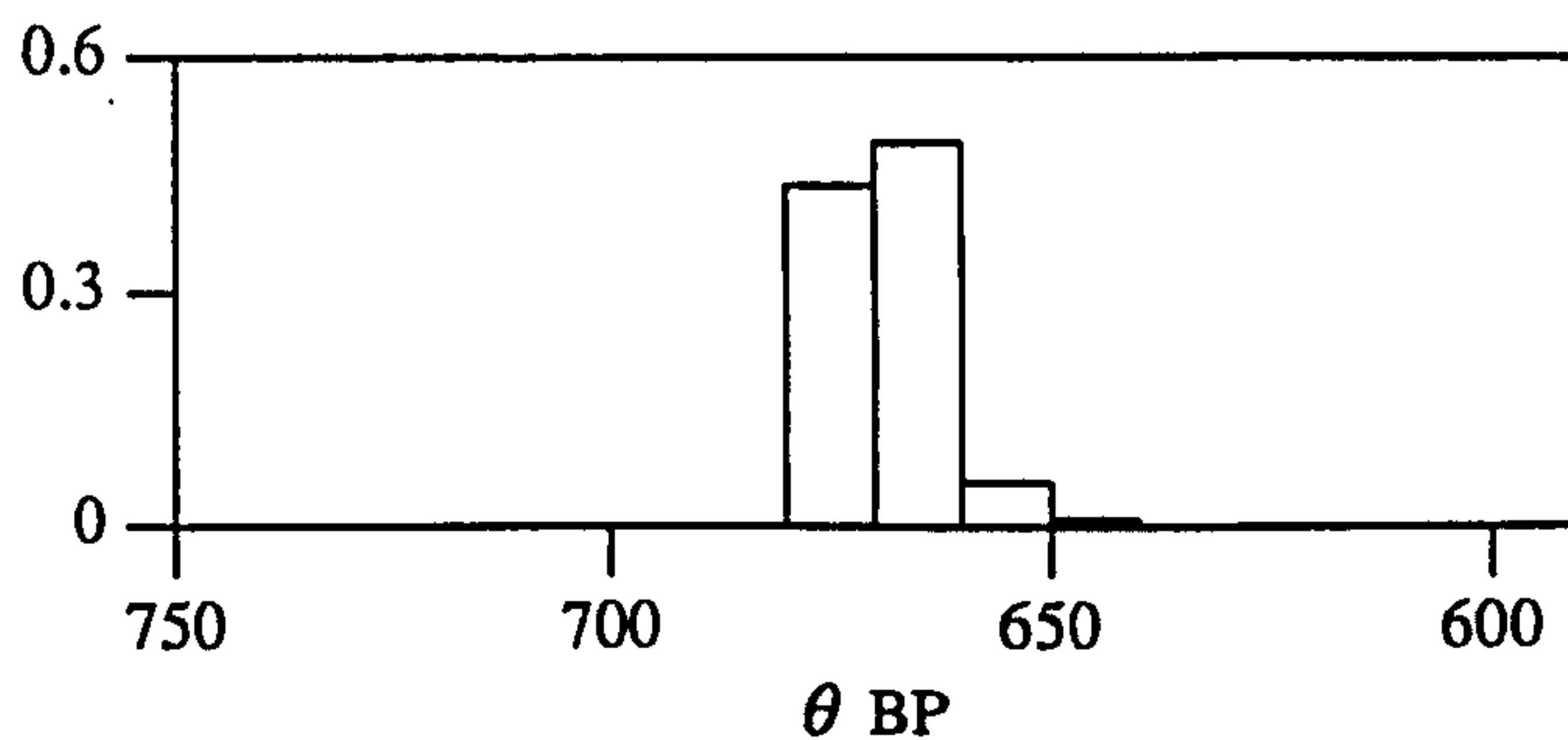


Figure 4.8
Histogram for the posterior distribution of θ ,
the 'Shroud of Turin', considering the variance in the
calibration curve.

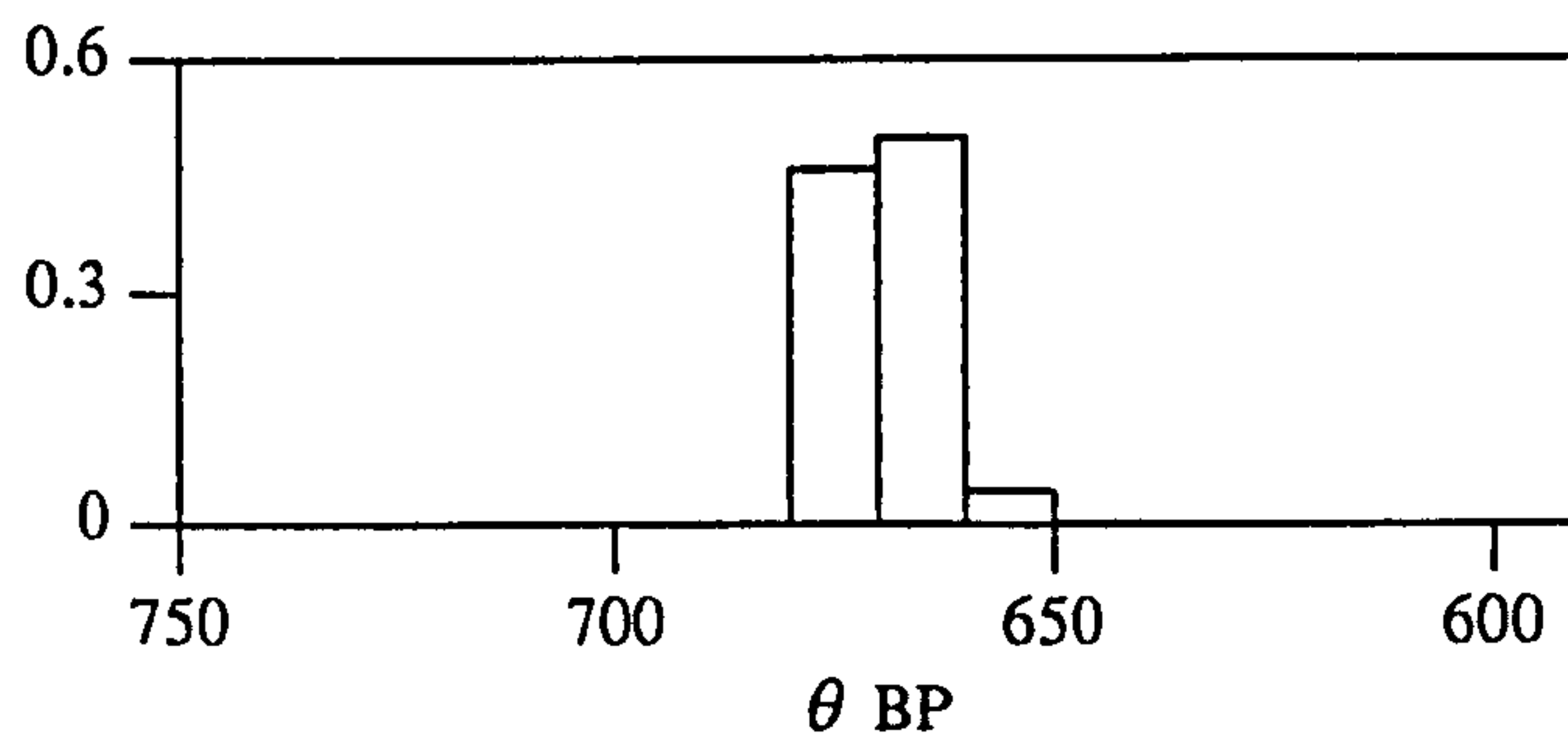


Figure 4.9
Histogram for the posterior distribution of θ ,
the 'Shroud of Turin', not considering the variance in the
calibration curve.

evidence shows that the organic materials contained in the Shroud died in this period. Of course we expect the Shroud's manufacture and possibly its first exhibition to have occurred soon after its organic materials died. Thus it seems likely that it was made some time between 1300 and 1350 AD, just as concluded in Damon *et al.* (1989b).

4.4.2 Comment

Of course, as we have mentioned before, there is the question of, when can a set of radiocarbon determinations be considered a set of replications? Strictly we

must have definite evidence telling us that the organic material contained in *all* the objects radiocarbon dated died in the same year. Thus, if we state our model as in Equation 4.6, θ would represent such a calendar year. Cases of this can be, for example, two parts of a bone, bones of two people known to have been killed in the same battle, seeds known to belong to the same harvest or, as above, three pieces of a shroud. As a modelling approximation we could consider a set of determinations to form a set of replications if the associated calendar years are known to be very similar. This could be, for example, two different tree rings in a small piece of timber. However, unless we have definitive evidence about this, we do not recommend considering any other sets of radiocarbon determinations as a set of replications.

If we have two radiocarbon determinations $y_1 \pm \sigma_1$ and $y_2 \pm \sigma_2$ arising from two different objects, and we state our model as in Equation 4.6, we will be assuming that the organic materials in both of the objects died in precisely the same calendar year (θ). If both 'objects' cannot be considered to have the same (or very similar) associated calendar years, it is difficult to have a correct interpretation for the meaning of θ . What would θ represent if we cannot assume that both objects are associated with the same calendar year?

As discussed in Section 2.2.2, where we reviewed other approaches, there is the idea that if a set of radiocarbon determinations are close together (measured with some statistical test) then the associated calendar years for the determinations will be close together. Thus the set of radiocarbon determinations can be modelled as a set of replications. We discussed also in Section 2.2.2 the evident inadequacies of such an approach because the radiocarbon calibration curve is non-monotonic. Thus, even if the radiocarbon determinations are judged to be similar, or with 'no significant difference', we cannot assume them to be associated with the same calendar year, unless we have other archaeological information

available. Moreover, there is still the possibility of one or more outliers in a sample and we must consider how they would affect the posterior distribution of θ . We leave the discussion of this problem until Chapter 6.

Chapter 5

Examples

5.1 Introduction

In this Chapter we present a series of examples to illustrate how our statistical framework developed in the previous Chapter, can be applied to the analysis and interpretation of sets of radiocarbon determinations related to various dating problems. Here we concentrate on the technical side of the problem, studying the models that can be used in each situation and calculating the posterior distributions of interest. Although less emphasis is given to the archaeological soundness of the problem, in each case we comment upon the appropriateness of the models used and how they could be improved and/or generalised. Later, in Chapter 7, we will present a detailed analysis and interpretation of a set of radiocarbon determinations related to a complex dating problem. In this example more emphasis is given to the elicitation of the relevant archaeological information needed to build the corresponding statistical model, and the archaeological soundness of the resulting analysis.

By presenting a series of examples in this Chapter, we intend to tackle the statistical problems related to the interpretation of radiocarbon determinations reviewed in Chapter 2 (note that we have already considered the ‘replications’ problem in Section 4.4). In Section 5.2 we continue to present a novel approach to the ‘summarisation problem’, introduced in Chapter 4, thus providing an alternative to the existing approaches reviewed in Section 2.4.4. In Section 5.3 we present a series of examples related to the problem of ‘archaeological wiggle matching’ (AWM) reviewed in Section 2.4.5. New and, we believe, more realistic models than those considered by the original authors are formulated. The

'contemporaneous houses' dating problem reviewed in Section 2.4.7 is analysed in Section 5.4 in the light of our framework. A slight generalisation is given which includes our calibration method in the analysis. In Section 5.5 we present an interesting interpretation problem related to peat growth, where a quite complex model is used in the analysis. Finally, in Section 5.6 we present a general approach to tackle the problems of dating archaeological phases using radiocarbon determinations. Problems of this type have been studied before within the Bayesian framework (reviewed in Section 2.4.7) and here we develop an extension of that work.

5.2 The Chancay culture: a summarisation problem

We continue our analysis of the summarisation problem defined in Section 4.3.5 using a set of radiocarbon determinations arising from an archaeological research originating in Peru. Pazdur and Krzanowski (1991) report on a study of the pre-Hispanic Peruvian culture called 'Chancay'. There are 13 radiocarbon determinations from samples of wood taken from tombs associated with this culture (see Table 5.1).

| Sample id. | Det. | Sample id. | Det. |
|------------|---------------|------------|---------------|
| Gd-2819 | 520 ± 60 | Gd-5672 | 830 ± 50 |
| Gd-3396 | 430 ± 30 | Gd-5823 | 670 ± 40 |
| Gd-5304 | 460 ± 50 | Gd-5824 | 1140 ± 50 |
| Gd-5307 | 970 ± 50 | Gd-6189 | 1070 ± 60 |
| Gd-5309 | 910 ± 35 | Gd-6196 | 810 ± 70 |
| Gd-5310 | 1000 ± 50 | Gd-6197 | 900 ± 70 |
| Gd-5312 | 390 ± 45 | | |

Table 5.1
Radiocarbon determinations for the 'Chancay' culture, Peru,
Pazdur and Krzanowski (1991).

In this problem, each sample radiocarbon dated is clearly associated with the Chancay culture, since the samples were taken from tombs known to belong to the Chancay. The aim of the analysis of the radiocarbon determinations is to estimate the duration, in the calendar scale, of the Chancay culture which, in this case, is the archaeological *phenomenon* of interest. We will use the modelling approach developed in Section 4.3.5 to estimate the beginning, β , and the end, α , of this *phenomenon*. Pazdur and Krzanowski used a combination of heuristic techniques and computer software to reach their conclusions about the time-span occurrence of the Chancay culture on the basis of the radiocarbon determinations available. No formal statistical procedure was used by them. They proceed as follows.

Using a computer program for calibration, explained in Pazdur and Michczynska (1989), they calibrate each of the 13 determinations separately. From that they calculate the corresponding 95% HPD regions (in the calendar scale) for each determination. Their estimate given for the time-span of the Chancay culture is the maximum to the minimum calendar years covered by the HPD regions. That is, if H_j is the 95% HPD region corresponding to determination y_j , then, letting $A = \bigcup_{j=1}^{13} H_j$, their estimate for the time-span is $\max(A)$ to $\min(A)$ years BP. (Note that it might be the case that $(\min(A), \max(A))$ is not equal to A ; this issue is not discussed by Pazdur and Krzanowski.) Evidently, this procedure will depend strongly on the probability chosen for the HPD sets. However, no measure is given for the accuracy of the estimates obtained and, since no statistical model is introduced, it is not clear how such a measure could be obtained. Moreover, no provision is given to allow for further archaeological information to be included in the analysis. The approach we present below avoids these problems.

In this example we have $y_1 \pm \sigma_1, y_2 \pm \sigma_2, \dots, y_{13} \pm \sigma_{13}$ associated with the (unknown) calendar years $\theta_1, \theta_2, \dots, \theta_{13}$. As explained before, our modelling approach to the summarisation problem takes the parameter ψ from our

framework to be $\psi = (\alpha, \beta)$ and assumes that

$$\theta_j \mid \alpha, \beta \sim U(\alpha, \beta)$$

for $j = 1, 2, \dots, 13$. That is, given α and β , θ_j has a uniform distribution within the beginning, β , and the end, α of the time-span of the Chancay culture. This represents the *a priori* knowledge that each sample is equally likely to have died at any time within the period of the Chancay culture. In this case, this seems to be an appropriate assumption since we have no prior information about the calendar dates for the wood samples taken from the tombs. We also assume that, given α and β , the θ_j 's are independent (that is, $f(\theta \mid \alpha, \beta) = \prod_{j=1}^m f(\theta_j \mid \alpha, \beta)$) and we take uniform prior distributions for both α and β , as in Equation 4.4. That is

$$\alpha \sim U(a_1, b_1) \text{ and } \beta \sim U(a_2, b_2).$$

Although the choice of a uniform distribution for both α and β might represent a strong constraint in this case, it is appropriate for the representation of the prior information available for α and β .

The Chancay culture is pre-Hispanic, and this means that the end of its time-span (α) should be before the invasion of Peru by Pizarro in the 16th century. In view of this we decided to set a_1 equal to 400 BP (1550 AD) as a late bound for the prior distribution of α . However, little is known about how early the Chancay culture began so we decided to set b_1 , a_2 and b_2 in a less informative way giving them extreme values ($b_1 = 600$, $a_2 = 601$ and $b_2 = 3000$). On the whole, the prior information used is vague (apart from the value of a_1), meaning that the posterior distributions of α and β will be based more on the radiocarbon determinations, and less on archaeological considerations.

The full conditionals for the parameters in our model are

$$f(\theta_j \mid y, \theta_{-j}, \alpha, \beta) = f(\theta_j \mid y_j, \alpha, \beta) \propto$$

$$\frac{1}{\omega_j(\theta_j)} \exp \left\{ -\frac{1}{2} \left(\frac{(y_j - \mu(\theta_j))^2}{\omega_j^2(\theta_j)} \right) \right\} I(\theta_j)_{(\alpha, \beta)},$$

$$f(\alpha \mid \theta, \beta) \propto (\beta - \alpha)^{-m} I(\alpha)_{(a_1, b)}$$

and

$$f(\beta \mid \theta, \alpha) \propto (\beta - \alpha)^{-m} I(\beta)_{(a, b_2)},$$

where $b = \min(\min(\theta), b_1)$ and $a = \max(\max(\theta), a_2)$.

For practical reasons and since the standard deviations are large (> 30), in this case we use the calibration procedure explained in Chapter 1. Thus we assume that $\omega_j^2(\theta_j) \approx \sigma_j^2$. We use the Gibbs sampling technique to obtain the marginal posterior distributions of α and β , shown as histograms in Figure 5.1.

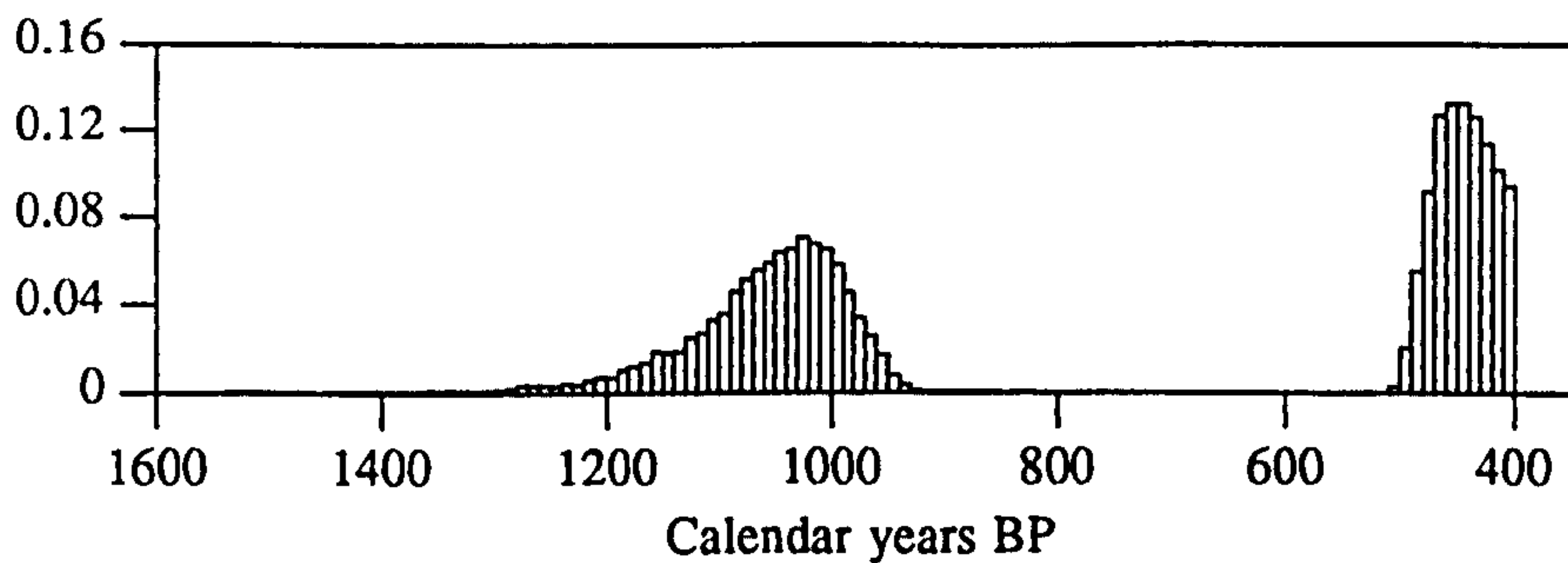


Figure 5.1

Histograms for the posterior distributions of β (left) and α (right),
Chancay culture's summarisation problem.

From the marginal distributions of α and β we see that, given the current sample, the 95% HPD region for α is approximately 480 to 400 BP, ie. 1470 to 1550 AD, and for β 1200 to 950 BP, ie. 850 to 1100 AD. The modes are at $\alpha = 440$ and $\beta = 1020$ BP, ie. 1510 and 930 AD respectively.

The final conclusion of Pazdur and Krzanowski for the time-span of the Chancay culture, on the basis of the radiocarbon datings performed, is from 1000

to 470 years BP (950 to 1480 AD). This in broad terms coincides with our range. However, we must mention that, in order to obtain their conclusion, Pazdur and Krzanowski removed from the sample the earliest and the latest radiocarbon determinations. They explained that the determinations removed belong to charcoal taken from tombs thought to be totally plundered, and for this reason their relation with the Chancay culture is '*incierta*' (uncertain). We agree that in some cases there are radiocarbon determinations that have an uncertain relationship with the archaeological *phenomena* under study. In such circumstances we could, from the beginning, exclude such radiocarbon determinations from the analysis (or have never sampled the material in the first place). It might be possible as well, using the techniques presented here, to consider such 'uncertain relationships' in the analysis via the specification of the prior distributions and avoid the removal of (very expensive) data. (For example, if y_j has an uncertain relationship with the Chancay culture, we could define its prior distribution as $f(\theta_j | \alpha, \beta) = wf_1(\theta_j | \alpha, \beta) + (1-w)f_2(\theta_j)$, where $f_1(\theta_j | \alpha, \beta)$ is the uniform distribution used before, w is a measure for that uncertainty and $f_2(\theta_j)$ is some alternative distribution.)

On the other hand, there are objects that on archaeological grounds do have a clear relationship with the *phenomena*, but whose corresponding radiocarbon determinations do not fit with our prior knowledge nor with the rest of the radiocarbon determinations performed. This leads to the problem of outliers in the sample, a topic that will be discussed in detail in Chapter 6 where we will reanalyse the Chancay data from that view point.

5.3 Floating chronologies

Suppose that we have a set of radiocarbon determinations $y_1 \pm \sigma_1, y_2 \pm \sigma_2, \dots, y_m \pm \sigma_m$ associated with the (unknown) calendar years

$\theta_1, \theta_2, \dots, \theta_m$. We have a 'floating chronology' when we have prior information about the relative dates $\theta_j - \theta_{j-1}$ for $j = 2, 3, \dots, m$. In archaeology there are two common examples of this. The first one is when the radiocarbon determinations arise from tree-rings and we have counted the number of rings in between samples. This is generally known as a 'tree-ring chronology'. In this case we can assume that we have, with no loss of generality,

$$\theta_j - \theta_{j-1} = l_j \text{ for } j = 2, 3, \dots, m$$

with $l_j > 0$ known.

The second case is when the radiocarbon determinations arise from samples taken from a 'stratigraphical sequence', for which we know that the associated calendar years must have a specific ordering and thus we have some knowledge about the relative dates $\theta_j - \theta_{j-1}$. Here we can assume, with no loss of generality, that we know (*a priori*)

$$\theta_j - \theta_{j-1} > 0 \text{ for } j = 2, 3, \dots, m.$$

Furthermore, given the specific problem at hand we might have further prior information about the relative dates $\theta_j - \theta_{j-1}$, for example, maximum and minimum time spans or a distribution for their length, etc. Note that in both cases we are assuming that

$$\theta_m > \theta_{m-1} > \dots > \theta_2 > \theta_1.$$

Problems of this type have been studied by other authors and some techniques have been proposed to tackle special examples (reviewed in Sections 2.3.4 and 2.4.5). To refer to those techniques some authors use the term 'archaeological wiggle matching' (AWM). Our approach considers the general case and is as follows.

Since the former case of floating chronologies explained above is a special case of the latter, we develop the general methodology with the latter case in mind. According to our framework we need to define ψ and then define $f(\psi_i | \psi_{-i})$ and $f(\theta_j | \theta_{-j}, \psi)$. In this problem we will be mainly interested in the posterior distributions of the θ_j 's, since the problems of floating chronologies are completely determined in terms of those parameters. We then consider ψ to be non-random and concentrate our modelling in defining $f(\theta_j | \theta_{-j})$.

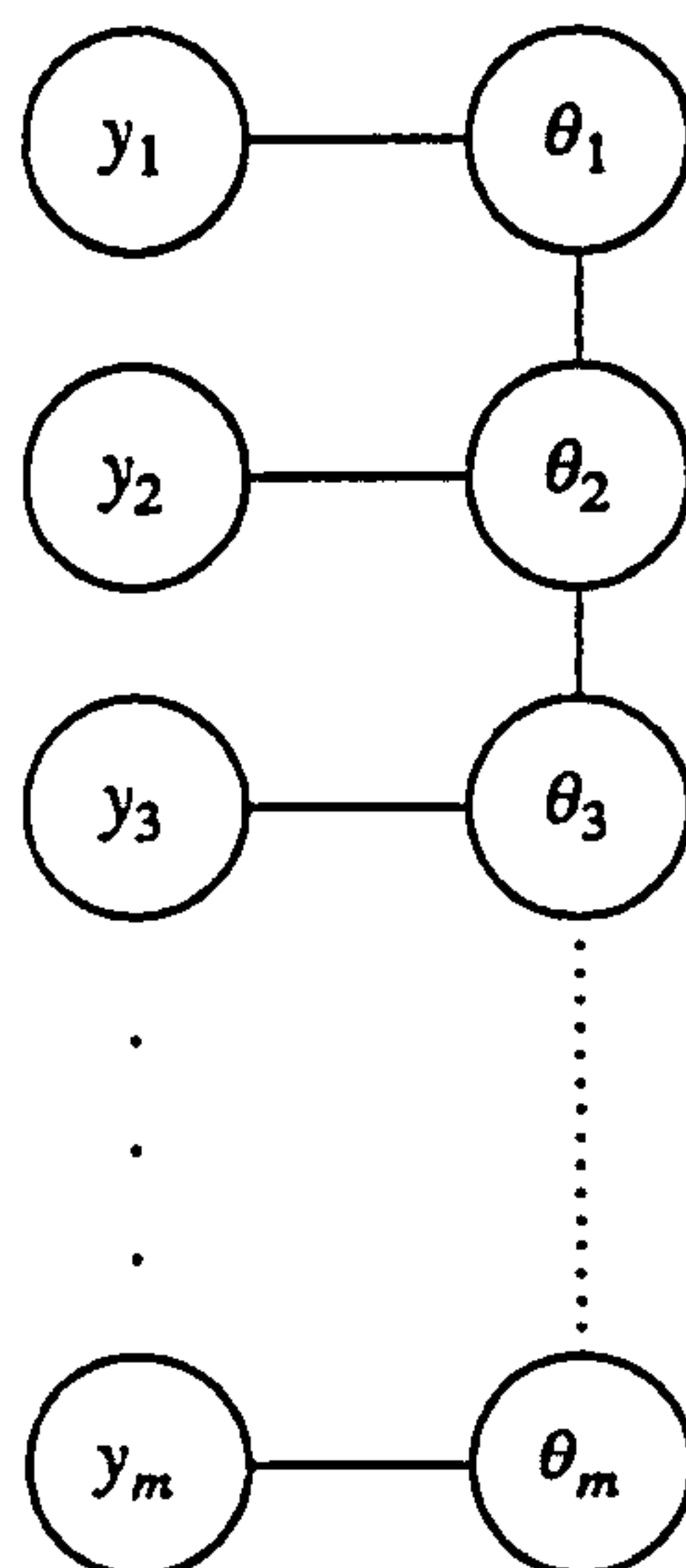


Figure 5.2
Hierarchy diagram for the general case
of a floating chronology.

Since we have assumed that $\theta_m > \theta_{m-1} > \dots > \theta_2 > \theta_1$, it is clear that $f(\theta_j | \theta_{-j}, \psi)$ should have support contained in the interval $(\theta_{j-1}, \theta_{j+1})$. From this it seems reasonable to assume that $f(\theta_j | \theta_{-j}) = f(\theta_j | \theta_{j-1}, \theta_{j+1})$. That is, *a priori*, given θ_{-j} , θ_j depends only upon θ_{j-1} and θ_{j+1} (indeed, for θ_1 and θ_m we have $f(\theta_1 | \theta_{-1}) = f(\theta_1 | \theta_2)$ and $f(\theta_m | \theta_{-m}) = f(\theta_m | \theta_{m-1})$). The corresponding hierarchy diagram for this problem is presented in Figure 5.2.

The key point in this problem is that we have prior information about the relative dates $\theta_j - \theta_{j-1}$, which ought to be included in the analysis. We will include this information via the specification of $f(\theta_j | \theta_{j-1}, \theta_{j+1})$. A sufficiently general procedure for doing this (that works correctly in our examples) is to assume that, *a priori*, the relative dates $\theta_j - \theta_{j-1}$ are independent. Therefore, given that the prior information about $\theta_j - \theta_{j-1}$ is defined with the density (or mass) function g_j , we have

$$f(\theta_j | \theta_{-j}) = f(\theta_j | \theta_{j-1}, \theta_{j+1}) \propto g_j(\theta_j - \theta_{j-1}) g_{j+1}(\theta_{j+1} - \theta_j)$$

for $j = 2, 3, \dots, m-1$ and,

$$f(\theta_1 | \theta_2) = g_2(\theta_2 - \theta_1) \text{ and } f(\theta_m | \theta_{m-1}) = g_m(\theta_m - \theta_{m-1}).$$

We use the distribution $f(y | \theta)$ as in Equation 4.1, that is

$$y_j | \theta_j \sim N(\mu(\theta_j), \omega_j^2(\theta_j))$$

(where $\omega_j^2(\theta_j) = \sigma_j^2 + \sigma^2(\theta_j)$) and we assume that, given the θ_j 's, the y_j 's are independent and therefore

$$f(y | \theta) = \prod_{j=1}^m f(y_j | \theta_j).$$

The full conditionals are

$$f(\theta_j | y, \theta_{-j}) = f(\theta_j | y, \theta_{j-1}, \theta_{j+1}) \propto \frac{1}{\omega_j(\theta_j)} \exp \left\{ -\frac{1}{2} \left(\frac{(y_j - \mu(\theta_j))^2}{\omega_j^2(\theta_j)} \right) \right\} g_j(\theta_j - \theta_{j-1}) g_{j+1}(\theta_{j+1} - \theta_j). \quad (5.1)$$

Then, different cases of floating chronologies will be modelled using different functions g_j 's. This will be illustrated in the next Sections by the use of two examples.

5.3.1 A tree-ring chronology

We now turn to the case of 'tree-ring chronologies'. As stated before we have

$$\theta_j - \theta_{j-1} = l_j$$

with $l_j > 0$ known for $j = 2, 3, \dots, m$, with the corresponding hierarchy diagram shown in Figure 5.3. Since $\theta_j = \theta_{j-1} + l_j$, we represent in our hierarchy diagrams deterministic relationships using a double arrowed linkage, as seen in Figure 5.3.

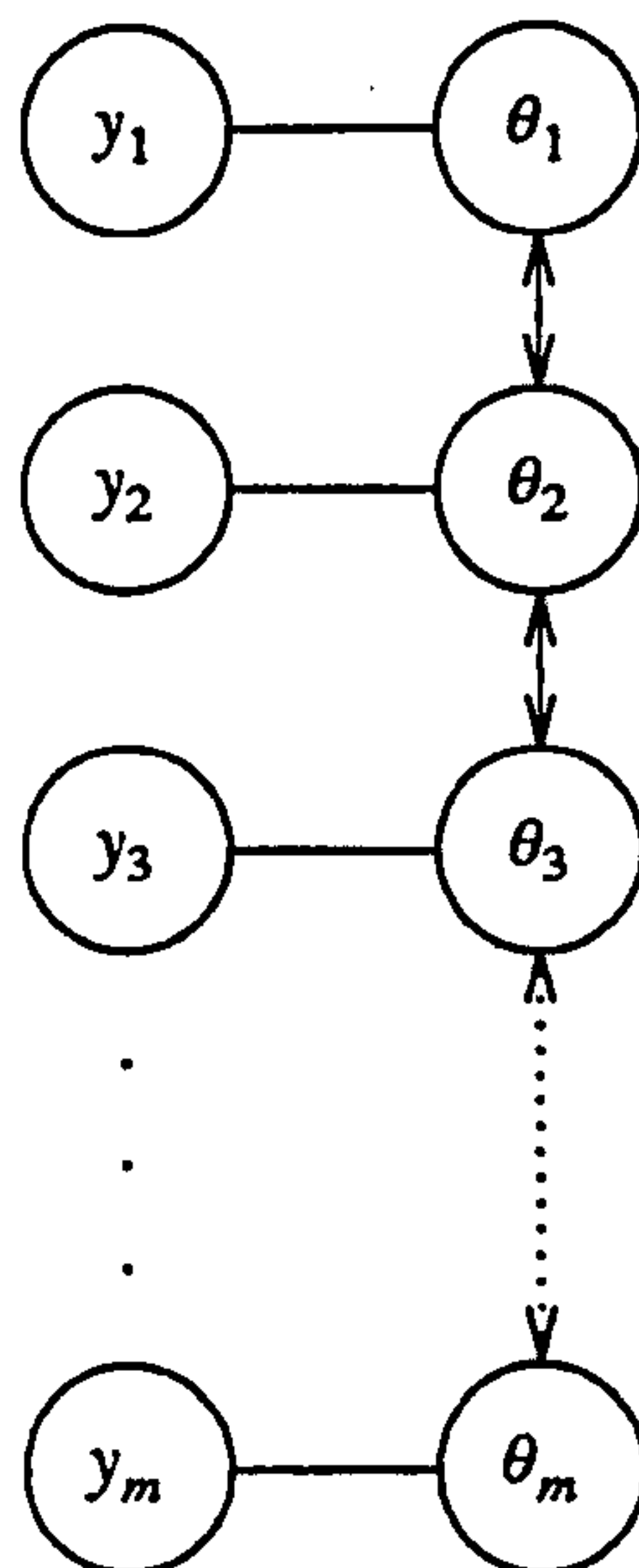


Figure 5.3

Hierarchy diagram for a tree-ring floating chronology.
Double arrowed linkages mean deterministic dependencies.

Since the l_j 's are known, in this case $\theta_j - \theta_{j-1} = l_j$ with probability one and, therefore, using the above functions g_j 's we have

$$g_j(x) = \begin{cases} 1 & \text{if } x = l_j \\ 0 & \text{if } x \neq l_j. \end{cases}$$

Here it is only necessary to calculate the distribution of θ_1 (the 'base' of the chronology) and the rest of the θ_j 's can then be calculated since $\theta_j = \theta_1 + \sum_{k=2}^j l_k$.

The distribution is given by

$$f(\theta_1 | y) \propto \left\{ \prod_{j=1}^m \frac{1}{\omega_j(\theta_j)} \right\} \exp \left\{ -\frac{1}{2} \sum_{j=1}^m \frac{(y_j - \mu(\theta_j))^2}{\omega_j^2(\theta_j)} \right\}$$

where $\theta_j = \theta_1 + \sum_{k=2}^j l_k$. In this case we can use a numerical integration procedure to find $f(\theta_1 | y)$.

| j | Det. |
|-----|-----------|
| 1 | 1270 ± 15 |
| 2 | 1292 ± 15 |
| 3 | 1268 ± 15 |
| 4 | 1263 ± 15 |
| 5 | 1290 ± 15 |

Table 5.2

Radiocarbon determinations from 'Ardnagross Horizontal Mill',
Pearson (1986). Tree-ring chronology, 20 year gaps between samples,
determination 1 (youngest) to determination 5 (oldest).

To illustrate our methodology we use an example reported by Pearson (1986). The problem is to date a floating tree-ring chronology using radiocarbon. Five consecutive blocks of 20 tree-rings taken from a piece of wood belonging to the 'Ardnagross Horizontal Mill' were radiocarbon dated. The piece of wood could not be dendrochronologically dated (for reasons unspecified by Pearson 1986) and so radiocarbon was used to date the wood.

Here we have $m = 5$ and $\theta_j - \theta_{j-1} = l_j = 20$. The determinations are given in Table 5.2. Unfortunately, the laboratory identifications are not given in Pearson (1986). Using numerical integration we obtained the posterior distribution for the base of the chronology θ_1 shown in Figure 5.4.

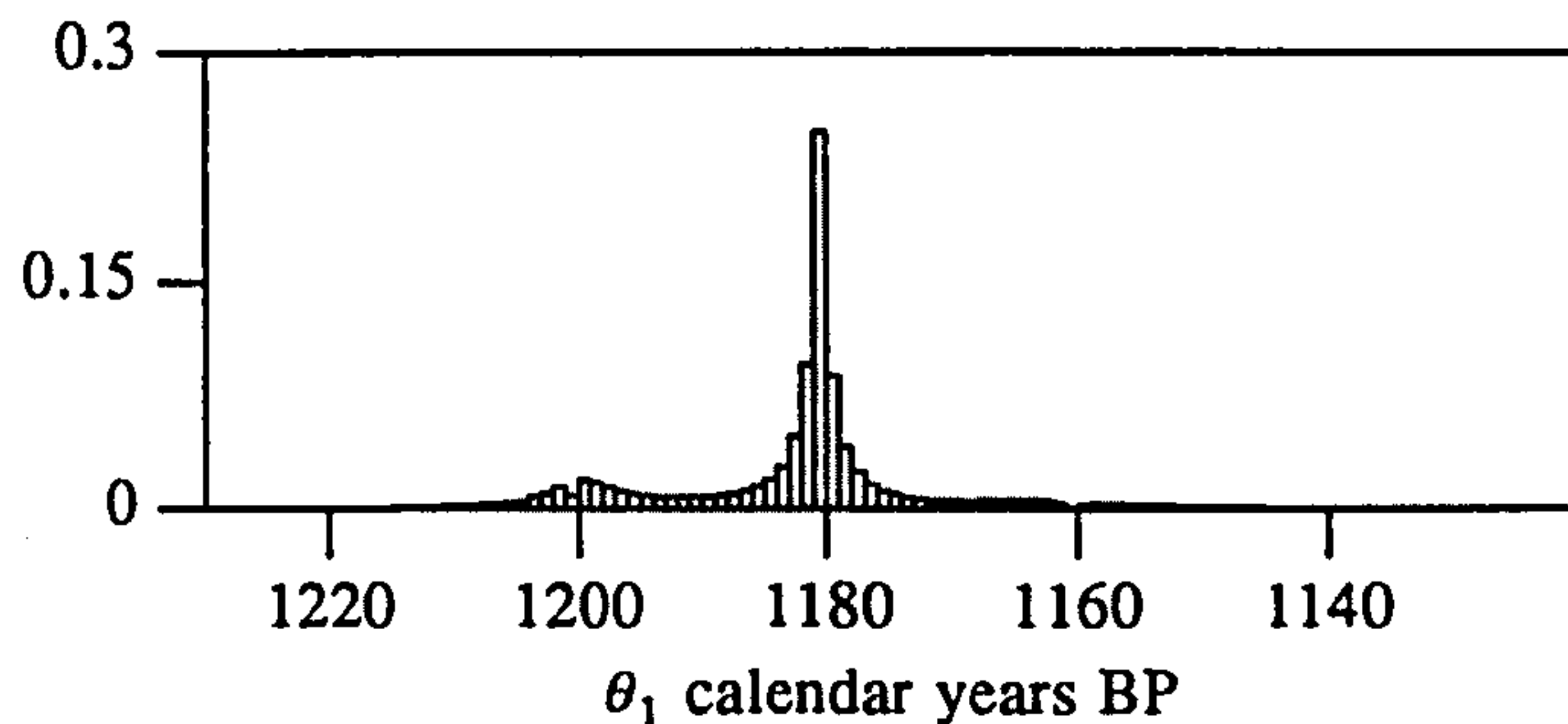


Figure 5.4
Posterior distribution for the base of the
floating tree-ring chronology 'Ardnagross Horizontal Mill'.
The MAP estimator is 1180 BP.

Pearson (1986), using a technique based on least squares (reviewed in Section 2.4.5), which consists of finding θ_1 such that

$$S(\theta_1) = \sum_{j=1}^m \frac{1}{\sigma_j^2} \left\{ y_j - \mu \left(\theta_1 + \sum_{k=2}^j l_k \right) \right\}^2$$

has a minimum, found the 'best' match for the base of the chronology. This minimum occurs when $\theta_1 = 1185$. Using our methodology the maximum *a posteriori* (MAP) estimator for the posterior distribution of θ_1 is 1180 (see Figure 5.4). The difference between the 'best' point estimates in both cases arises from the fact that in our analysis (since the variances σ_j^2 's are quite small) we are considering the variance $\sigma^2(\theta)$ in the calibration curve. If we had neglected the variance $\sigma^2(\theta)$ we would have that $f(\theta_1 | y) \propto \exp(-0.5 S(\theta_1))$ and thus the MAP estimator will coincide with the minimum of $S(\theta_1)$. When the variance $\sigma^2(\theta)$ is included in the analysis this is not necessarily the case.

To illustrate the above in Figure 5.5 we have plotted the radiocarbon determinations used in this example, along with the relevant part of the calibration curve $\mu(\theta)$ and $\mu(\theta) - \sigma(\theta)$ and $\mu(\theta) + \sigma(\theta)$. Note that at $\theta_1 = 1180$ the radiocarbon determinations coincide with knots of the calibration curve (since the

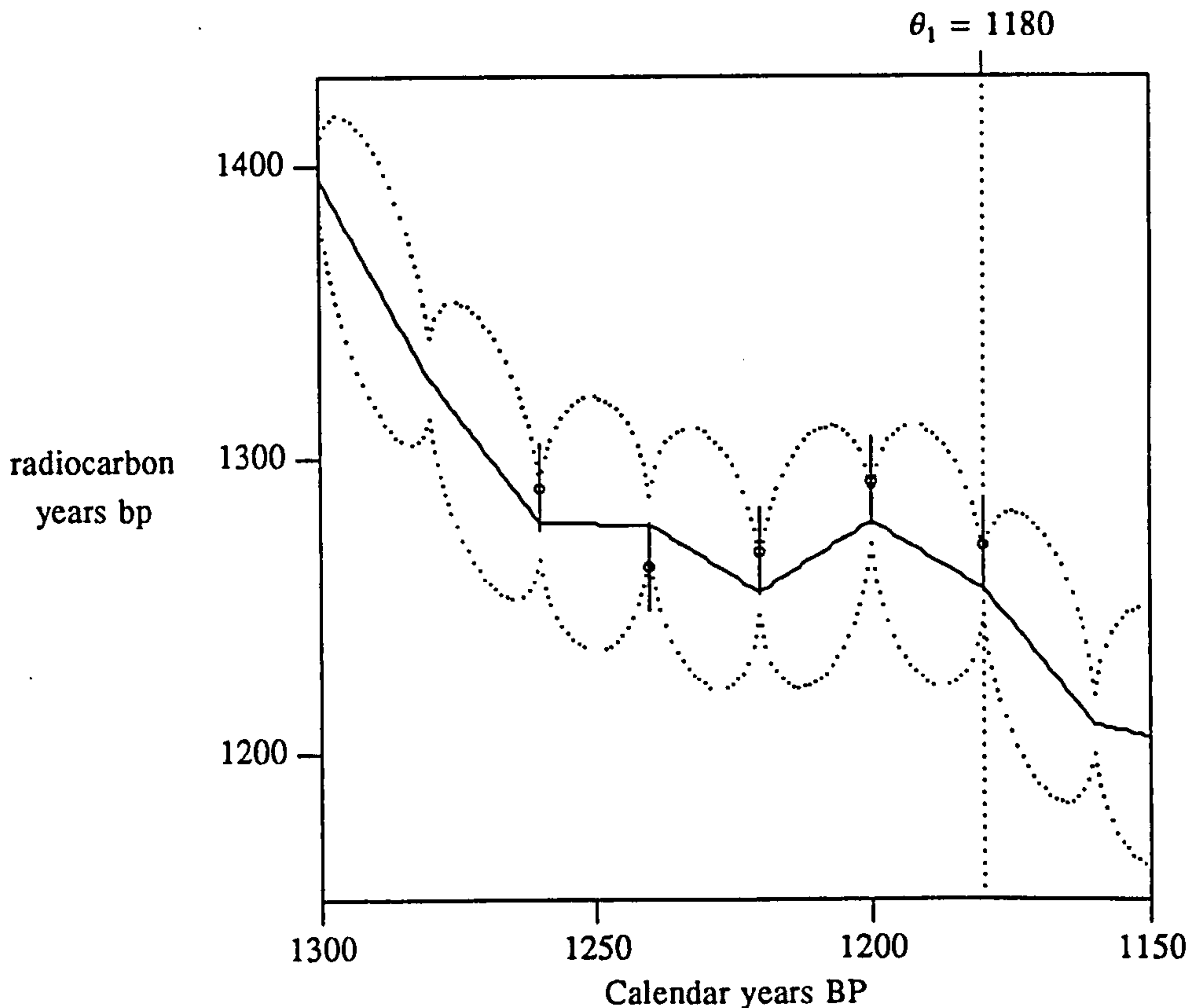


Figure 5.5

Wiggle match, using $\mu(\theta)$ and $\sigma(\theta)$, for the floating tree-ring chronology reported in Pearson (1986), for the MAP estimator of θ_1 .

The circles represent the y_j 's and the lines are from $y_j - \sigma_j$ to $y_j + \sigma_j$.

radiocarbon determinations are spaced every 20 years). At the knots the variance $\sigma^2(\theta)$ is smallest since we have observed the calibration curve there and, therefore, our method prefers, as a point estimate, $\theta_1 = 1180$. If we neglect the variance $\sigma^2(\theta)$, the resulting technique makes no distinction between points on the calibration curve (knots or outside knots) and therefore the best point estimate will be the minimum of the squares $S(\theta_1)$.

As seen above, the method of least squares and our Bayesian method are in agreement, at least for the case of point estimates. But: how can we assess the precision of the least squares match? How sure are we about such a match? We believe that, in general, a probability distribution for the base of the chronology

θ_1 gives a much more reliable and informative measure for dating floating tree-ring chronologies using radiocarbon. (For example, using our method, we can say that the 95% HPD region for θ_1 is 1200 to 1160 BP.)

5.3.2 A floating chronology

To study the problems of archaeological wiggle matching (AWM), in the light of our methodology, we analyse a data set used by Manning and Weninger (1992), where the technique of AWM was fully illustrated. Surprisingly, the ‘floating chronologies’ used in that paper were not formed of a series of radiocarbon determinations for which their associated calendar years were chronologically ordered. Rather, the ‘floating chronologies’ consisted of a series of archaeological phases which were stratigraphically ordered, with several radiocarbon determinations related to each phase (we discuss problems of this kind in Sections 2.4.7 and 5.6).

The actual approach followed by Manning and Weninger was to calculate the ‘pooled means’ of each group of determinations and use such pooled means as ‘radiocarbon determinations’ associated with a single calendar year. Using these averaged ‘determinations’ they then applied their techniques for AWM.

In Weninger (1986) and Manning and Weninger (1992) the underlying assumptions involved in the usage of pooled means was not discussed. We have discussed the potential inadequacies of such an approach in Sections 2.4.3 and 4.4. Nevertheless, and only for the sake of comparing techniques, we study the ‘floating chronology’ found in Manning and Weninger (1992), where the ‘determinations’ quoted are actually ‘pooled means’ (the real data is not published in that paper).

| Schicht | <i>j</i> | pooled mean | Schicht | <i>j</i> | pooled mean |
|---------|----------|-------------|---------|----------|-------------|
| 6 | 1 | 2825±67 | 12 | 7 | 2982±15 |
| 8 | 3 | 2952±33 | 13 | 8 | 2949±29 |
| 9 | 4 | 2980±50 | 14 | 9 | 3121±24 |
| 10 | 5 | 2899±37 | 16 | 11 | 3136±17 |
| 11 | 6 | 2965±45 | 17 | 12 | 3180±55 |

Table 5.3

Pooled means from a stratified sequence of radiocarbon determinations
for the Late Bronze Age at Kastanas, Greece.
Taken from Manning and Weninger (1992).

The archaeological problem considered is that of dating a series of chronologically ordered 'Schichten' (layers) from Kastanas, Greece. Radiocarbon determinations were obtained for 11 Schichten and the corresponding pooled means were calculated. On archaeological grounds the pooled mean for Schicht 18 was rejected as an outlier and thus we only use the remaining 10 pooled means. These pooled means are presented in Table 5.3.

The key feature in the problem is that successive numbered Schichten are known to be chronologically ordered. The AWM technique used by Manning and Weninger (1992) allows only the possibility of a fixed length of time between Schichten. That is, they assumed, perhaps unrealistically, that the time elapsed between successive Schichten was equal. In other words, if $\theta_1, \theta_3, \theta_4, \dots, \theta_9, \theta_{11}, \theta_{12}$ are the associated calendar years for the pooled means in each Schicht, then

$$\theta_3 - \theta_1 = 2l$$

$$\theta_j - \theta_{j-1} = l \quad j = 4, 5, \dots, 9, 12$$

$$\theta_{11} - \theta_9 = 2l$$

for some fixed l . Then they fixed l at multiples of 5, for 10 to 60 years

($l = 10, 15, \dots, 60$) and concluded that the best match was when $l = 35$. The corresponding value for θ_j given by their method at each value of l was then averaged and hence they could calculate a 'standard deviation'. These 'averaged' dates, together with the standard deviations, are given in Table 5.4 column (i).

| Schicht | (i) BP | (ii) BP | (iii) yrs | (iv) BP | (v) BP | (vi) yrs | (vii) BP |
|---------|-----------|------------|--------------|------------|-----------|-------------|-------------|
| 6 | 2995±75 | 2956 | 48 | 2860-3050 | 2970 | 38 | 2900-3040 |
| 8 | 3060±50 | 3038 | 38 | 2960-3100 | 3043 | 32 | 2980-3100 |
| 9 | 3090±40 | 3079 | 32 | 3020-3130 | 3077 | 30 | 3020-3130 |
| 10 | 3125±25 | 3120 | 26 | 3060-3100 | 3111 | 29 | 3050-3160 |
| 11 | 3160±20 | 3160 | 24 | 3120-3200 | 3147 | 27 | 3100-3200 |
| 12 | 3190±25 | 3201 | 17 | 3160-3250 | 3184 | 24 | 3140-3230 |
| 13 | 3225±30 | 3242 | 16 | 3200-3270 | 3223 | 23 | 3180-3260 |
| 14 | 3255±35 | 3283 | 17 | 3250-3320 | 3270 | 20 | 3220-3300 |
| 16 | 3320±50 | 3365 | 24 | 3325-3410 | 3352 | 22 | 3320-3400 |
| 17 | 3350±65 | 3405 | 23 | 3360-3460 | 3387 | 25 | 3340-3440 |

Table 5.4

Column (i): average best fit found by Manning and Weninger (1992)
for the wiggle matching of ten Schichten from Kastanas, Greece.

Columns (ii) to (vii): posterior means, standard deviations and 95% HPD regions
in our analysis for priors $l = 10, 15, \dots, 60$ and $l_i \sim G(5, 7)$.

We reanalyse this data using our Bayesian methodology as follows. Firstly we assume that the time elapsed between successive Schichten is equal. We maintain the assumption that $l = 10, 15, \dots, 60$ and assume that, *a priori*, any of those values is equally likely. Thus, for $j = 4, 5, \dots, 9, 12$ we have

$$g_j(x) = 1/11 \text{ if } x = 10, 15, \dots, 60,$$

and for $j = 3$ and $j = 11$

$$g_j(x) = 1/11 \text{ if } x = 20, 40, \dots, 120.$$

In this case we have the likelihood

$$L(y \mid \theta_1, l) \propto \prod_{\substack{j=1 \\ j \neq 2,10}}^{12} \frac{1}{\omega_j(\theta_j)} \exp \left\{ -\frac{1}{2} \frac{(y_j - \mu(\theta_j))^2}{\omega_j^2(\theta_j)} \right\}$$

where $\theta_j = \theta_1 + (j-1)l$.

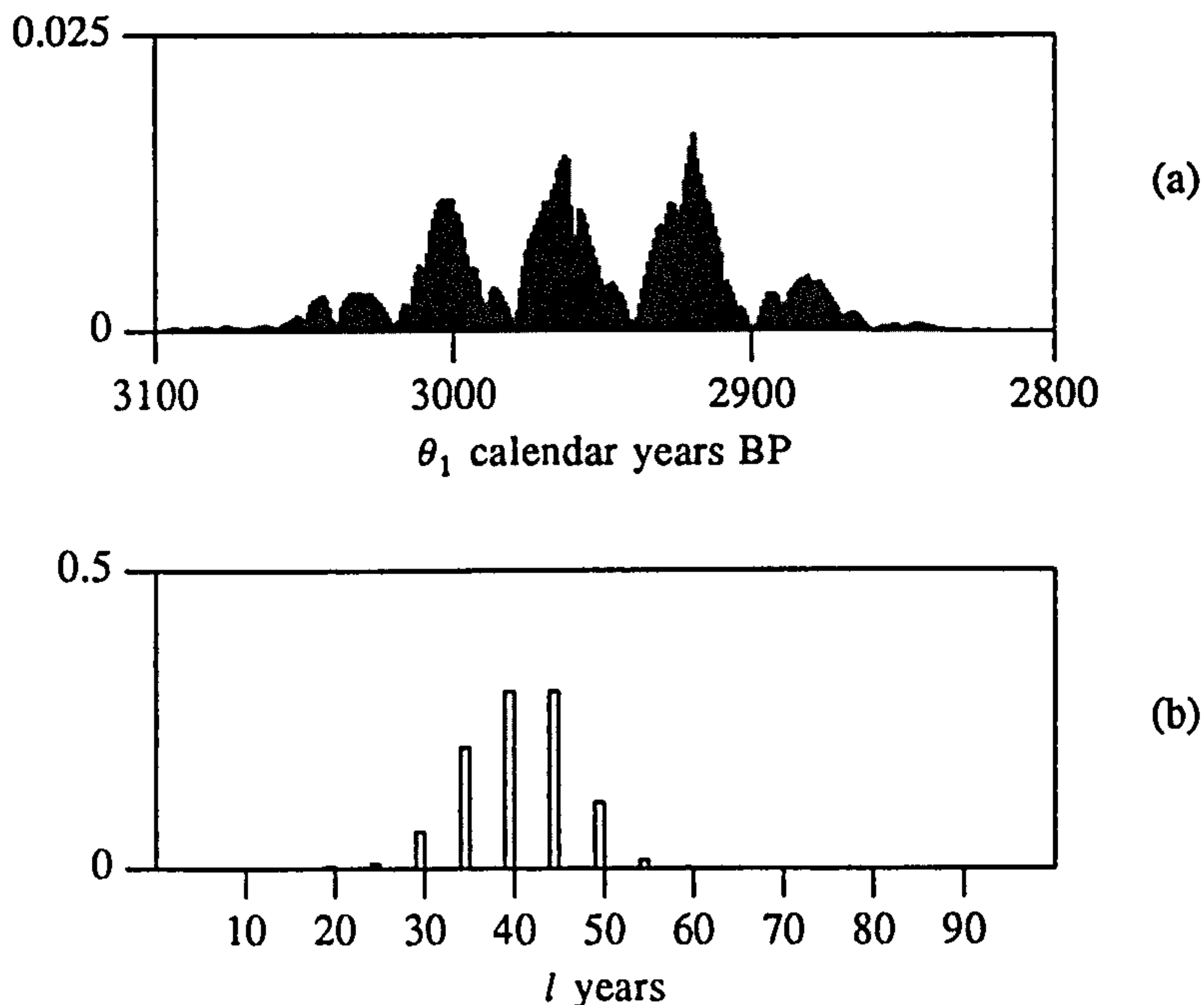


Figure 5.6
(a) Marginal posterior density for Schicht 6, θ_1 ,
and (b) marginal posterior mass function of l .

With the above definition for the functions g_j 's we calculate the posterior marginal distributions of θ_1 and l shown in Figure 5.6. Also, in Table 5.4 we present the posterior mean and standard deviation and the 95% HPD regions for each θ_j in columns (ii), (iii) and (iv) respectively. In this case the choice of a discrete prior distribution for l is reflected in a spiky distribution for θ_1 , as seen in Figure 5.6. Note that the most likely values for l are 40 and 45 with 35 next. This contrasts to Manning and Weninger's (1992) 'best' estimate for l of 35 years.

Secondly we relax two assumptions made by Manning and Weninger (1992).

These are:

- (i) The times between successive Schichten (l_j) are equal.
- (ii) The times between successive Schichten (l_j) are a multiple of 5 between 10 and 60 inclusive.

We do this by assuming the l_j 's to have a Gamma prior distribution. We took the parameters $a = 5$ and $b = 7$ to allow the l_j 's some chance of being less than 10 and greater than 60 years, and maintain the prior expectation equal to 35 years, as was the case in the former analysis. Here then, *a priori*, $P(l_j < 10) \approx 0.01$ and $P(l_j > 60) \approx 0.07$, see Figure 5.7.

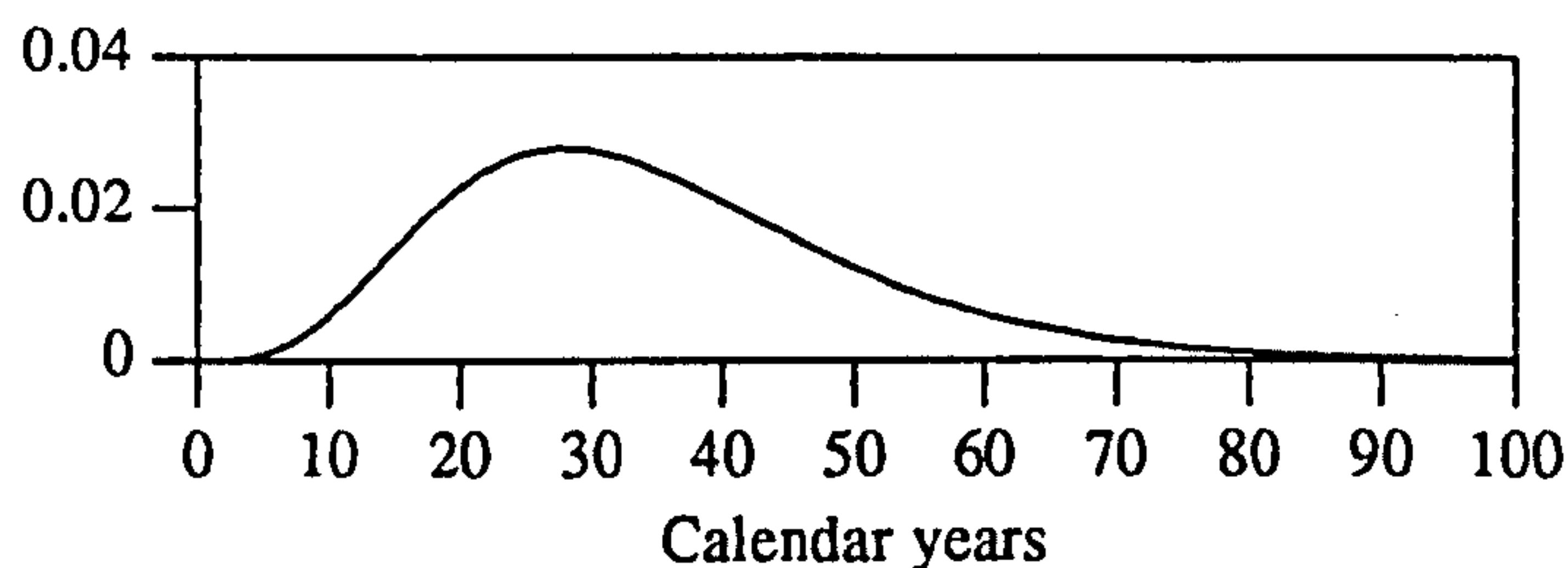


Figure 5.7

Prior density for the intervals $l_j \sim G(5, 7)$.

Here we allow the l_j 's to be less than 10 and greater than 60 since $P(l_j < 10) \approx 0.01$ and $P(l_j > 60) \approx 0.07$.

In terms of the functions g_j 's we have

$$g_j(x) = \frac{x^{a-1} \exp(-x/b)}{b^a \Gamma(a)} \quad \text{for } x > 0,$$

for $j = 4, 5, \dots, 9, 12$ and $a = 5$ and $b = 7$. Since $\theta_3 - \theta_1 = l_3$ and $\theta_{11} - \theta_9 = l_{11}$ we expect that, *a priori*, l_3 and l_{11} are twice the length of the rest of the l_j 's. Therefore, for $j = 3, 11$ we have

$$g_j(x) = \frac{x^{2a-1} \exp(-x/b)}{b^{2a} \Gamma(2a)} \quad \text{for } x > 0$$

(since $2z \sim G(2a, b)$ if $z \sim G(a, b)$).

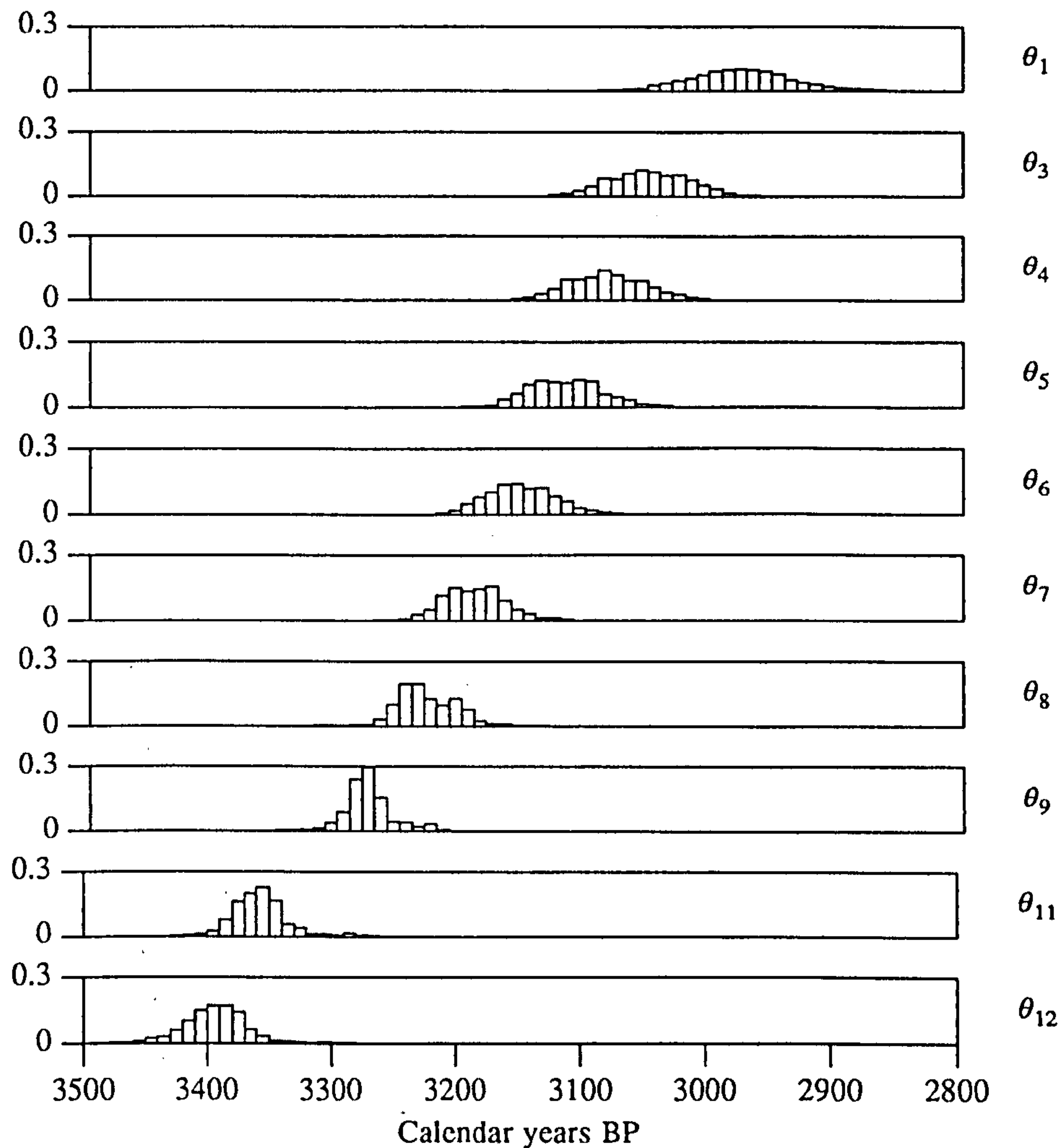


Figure 5.8
Marginal posterior densities for the θ_j 's assuming different and independent intervals l_j 's with prior $l_j \sim G(5, 7)$, for the ten Schichten from Kastanas, Greece.

Using the full conditionals given in Equation 5.1 we implemented the Gibbs sampler and, after checking for convergence, obtained the distributions for $\theta_j | y$ shown in Figure 5.8.

In columns (v) and (vi) of Table 5.4 we give the posterior mean and standard deviation of each θ_j and in column (vii) the 95% HPD regions for each θ_j . By comparing these columns with column (i) in the same Table, it can be seen then that the heuristic estimates of Manning and Weninger (1992) have some agreement with our results, though our standard deviations are smaller. In addition, by comparing columns (iv) and (vii) (the HPD regions for both analyses) we see that our results do not drastically change from the first to the second analysis performed. This tells us that results will be similar if we maintain similar priors.

We do not believe that the posterior distributions shown in Figure 5.8 have a definitive archaeological relevance since we have not analysed the consequences of transforming the original set of radiocarbon determinations into a series of pooled means. Moreover, it might be the case that the original set of radiocarbon determinations should be studied as an ‘archaeological phases’ problem and not viewed as a AWM problem (we will study archaeological phases problems in later Sections). This might lead to a more relevant means of analysis so giving the archaeologists more interpretable results.

5.4 Contemporaneous houses

In Section 2.4.7 we reviewed a problem first reported by Helskog and Schweder (1989). The problem is one of dating and checking for contemporaneous use of a set of houses from two adjacent sites in Norway. Radiocarbon determinations have been performed on charcoal belonging to 11 out of 23 houses. Unfortunately, the uncalibrated radiocarbon determinations were not given by Helskog and Schweder. We were unable to locate the determinations from the lists of results given in the journal *Radiocarbon*, and we did not get a response to a letter addressed to one of the authors. Therefore, our intention in this Section is to illustrate how this problem can be analysed within our statistical framework but

because of the lack of original data we are unable to compare results.

We study a slightly more general problem in which we have several radiocarbon determinations performed for each house (Helskog and Schweder 1989 studied the problem having only one determination per house). Suppose we have n houses and m_i determinations performed for house i . In this case we have a set of radiocarbon determinations formed by

$$y_{i,j} \pm \sigma_{i,j}; \text{ for } i = 1, 2, \dots, n \text{ and } j = 1, 2, \dots, m_i$$

and associated with the calendar years $\theta_{i,j}$'s. Let $u_{i,j}$ be the deposition time of sample j belonging to house i , τ_i the year of abandonment of house i and l_i the occupation longevity of house i .

Here we define the parameter ψ from our framework as $\psi = (\tau_1, l_1, \tau_2, l_2, \dots, \tau_n, l_n)$. We assume (as in Helskog and Schweder 1989) that the disposition time of all samples is similar to the associated calendar year for each determination and thus $\theta_{i,j} = u_{i,j}$. Using the model in Equation 4.1 we have

$$y_{i,j} \mid \theta_{i,j}, \tau_i, l_i \sim N(\mu(\theta_{i,j}), \omega_{i,j}^2(\theta_{i,j}))$$

(where $\omega_{i,j}^2(\theta_{i,j}) = \sigma_{i,j}^2 + \sigma^2(\theta_{i,j})$) and we assume that, given the θ_j 's, the $y_{i,j}$'s are independent.

Since the occupation period of house i is $\tau_i + l_i$ to τ_i (years BP), given both τ_i and l_i , the $\theta_{i,j}$'s should all belong to that period of time. That is, the organic matter contained in sample i,j should have died within the period of occupation of house i . Furthermore, assuming that the samples are equally likely to have died at any time within the occupation period of the houses we have

$$\theta_{i,j} \mid \tau_i, l_i \sim U(\tau_i, \tau_i + l_i).$$

Now, the deposition of one sample should not affect the deposition of another

sample and thus we assume that, given ψ , the $\theta_{i,j}$'s are independent. That is

$$f(\theta \mid \psi) = \prod_{i=1}^n \prod_{j=1}^{m_i} f(\theta_{i,j} \mid \tau_i, l_i).$$

Helskog and Schweder assumed that τ_i has a vague prior distribution and that l_j and τ_j were independent (that is, $f(\psi) = \prod_{i=1}^n f(l_i)f(\tau_i)$). Here we maintain these two assumptions. However, they assumed that l_i was fixed and equal for all houses. This latter assumption is relaxed here. An abbreviated hierarchy diagram for this model is presented in Figure 5.9

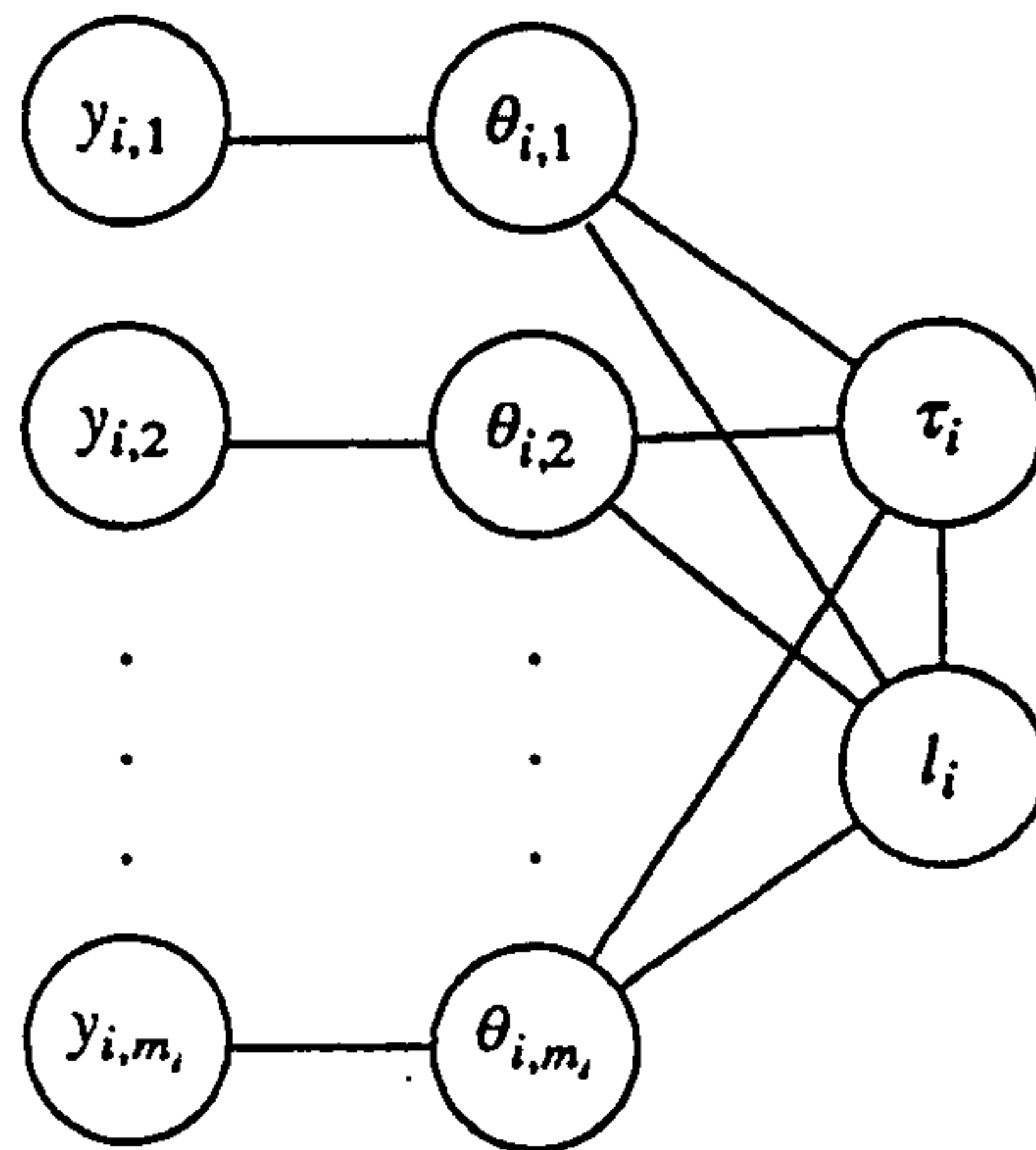


Figure 5.9

Abbreviated hierarchy diagram for the 'contemporaneous houses' problem.

The nodes above correspond to house i only. The complete diagram would be a series of the above diagram for $i = 1, 2, \dots, n$.

There are no other links than the ones shown.

From Equations 4.2 and 4.3, and the above considerations about $f(\psi)$ it is easy to see that the full conditionals are

$$f(\theta_{i,j} \mid y_{i,j}, \theta_{-i,j}, \tau_i, l_i) \propto \frac{1}{\omega_{i,j}(\theta_{i,j})} \exp \left\{ -\frac{1}{2} \frac{(y_{i,j} - \mu(\theta_{i,j}))^2}{\omega_{i,j}^2(\theta_{i,j})} \right\} I(\theta_{i,j})(\tau_i, \tau_i + l_i),$$

$$f(\tau_i \mid \theta_i, l_i) \propto \prod_{j=1}^{m_i} I(\theta_{i,j})(\tau_i, \tau_i + l_i)$$

and

$$f(l_i | \theta_i, \tau_i) \propto \frac{1}{l_i^m} \prod_{i=1}^n I(\theta_{i,j})_{(\tau_i, \tau_i+l_i)} f(l_i)$$

where $\theta_i = (\theta_{i,1}, \theta_{i,2}, \dots, \theta_{i,m_i})$. Note that the full conditional for τ_i is actually a uniform distribution on $\max(\theta_i) - l_i, \min(\theta_i)$ and that $I(\theta_{i,j})_{(\tau_i, \tau_i+l_i)}$, viewed as a function of l_j , is equal to $I(l_i)_{(\max(\theta_i)-\tau_i, \infty)}$. Therefore we have

$$\tau_i | \theta_i, l_i \sim U(\max(\theta_i) - l_i, \min(\theta_i))$$

and

$$f(l_i | \theta_i, \tau_i) \propto \frac{1}{l_i^m} I(l_i)_{(\max(\theta_i)-\tau_i, \infty)} f(l_i).$$

It is not difficult to sample from either of the above distributions (provided $f(l_j)$ has a simple form). Using a MCMC method we can then obtain $f(\psi | y)$. As stated in our framework, we should then base our conclusions using that distribution. As stated before, what we want to know is the number of contemporaneous houses at any time t . This can be calculated from $f(\psi | y)$. To see this we note that, given ψ , the number of contemporaneous houses at any time t is

$$N(t, \psi) = \sum_{i=1}^n I(t)_{(\tau_i, \tau_i+l_i)}.$$

Therefore, the expected number of contemporaneous houses at any time t , given y , is

$$N(t | y) = E[N(t, \psi) | y] = \int N(t, \psi) f(\psi | y) d\psi.$$

One can then implement a MCMC method to obtain $f(\psi | y)$ and after that calculate the above integral to obtain $N(t | y)$. This is not the most desirable way to proceed since the (MCMC) sampling process itself can give us an estimate for $N(t | y)$, by noticing that

$$\frac{1}{s} \sum_{i=1}^s N(t, \psi^{(i)}) \xrightarrow{d} N(t | y) \text{ as } s \rightarrow \infty$$

(see Section 4.3.4).

Note that $N(t | y)$ will not be a function of the l_i 's, as opposed to what Helskog and Schweder obtained. Still, there is the question about the assumption that $\theta_{i,j} = u_{i,j}$. That is, the time of deposition of the samples is equal to their associated calendar years. This assumption could be relaxed by letting $\theta_{i,j} = u_{i,j} + d_{i,j}$ where $d_{i,j}$ represents the time elapsed between the death of sample i,j and its deposition. Several options can be followed, from using a single $d = d_{i,j}$ for all determinations or giving the $d_{i,j}$'s a prior distribution. We do not explore this possibility any further.

5.5 Peat cores

Large parts of the earth are covered by peat and the study of peat cores is an important discipline that, among other things, helps us to understand environmental change. It is known that peat accumulates and decays through time and therefore an underlying relationship between peat depth and peat age can be expected. By performing chemical and botanical analyses of peat at successive depths one can obtain a picture of environmental change for the area under investigation. Therefore, it is crucial for environmental scientists to have a clear description for peat depth and age in order to understand the chronology of environmental change (see Clymo 1991 for a review of the subject).

In general, however, the relationship between peat depth and peat age is difficult to establish. It is here that radiocarbon dating plays an important rôle in peat core study, since it is used to help establish such a relationship. Peat cores are commonly extracted in cylindrical samples and sliced to generate a series of peat samples at various depths. Some of these samples are then radiocarbon

dated. It is then expected, using the corresponding determinations and some other theoretical considerations, to obtain a description for depth against age for the peat core under study.

Evidently there are a series of technical difficulties to be solved before a realistic description of peat depth against age can be achieved using radiocarbon dating. Here we have a series of radiocarbon determinations $y_1 \pm \sigma_1$, $y_2 \pm \sigma_2, \dots, y_m \pm \sigma_m$ arising from peat samples taken at successive depths x_1, x_2, \dots, x_m . Since peat age increases with depth, the associated calendar years for the radiocarbon determinations (the calendar years at which the organic materials contained in the peat samples died) $\theta_1, \theta_2, \dots, \theta_m$ should be ordered. That is (using years BP) $\theta_m > \theta_{m-1} > \dots > \theta_1$.

Furthermore, based upon theoretical considerations about peat growth below a certain fixed depth x_0 , and using the rate, ρ , at which mass is added (on an area basis) at depth x_0 and the proportional rate of mass decay, α ; the following model has been proposed (see Clymo 1991 for details). It is proposed that, for the dry mass M accumulated (on an area basis) below x_0 , we have

$$M = \frac{\rho}{\alpha} \{1 - \exp(-\alpha(\theta - \theta_0))\}, \quad (5.2)$$

where θ is the age of the peat below x_0 and θ_0 is the age of the peat at depth x_0 . Therefore the model relates the age of the peat with the cumulative dry mass which, in turn, is related to peat depth.

As can be seen from the above, several factors are involved in this problem. Ideally, all of the above considerations should be combined to obtain a description for peat depth and its age using the evidence of the radiocarbon determinations available for the core under study. Undoubtedly, this will involve radiocarbon calibration and further statistical modelling to obtain satisfactory results.

Given that the θ_j 's are known, *a priori*, to be chronologically ordered, wiggle matching techniques have been used to tackle this problem (see Clymo *et al.* 1990). However, results have not been totally satisfactory (we have discussed some of the inadequacies of wiggle matching techniques in Sections 2.4.5 and 5.4). The principal technical difficulty here is to consistently combine radiocarbon calibration, the theoretical considerations about peat growth explained above and the radiocarbon determinations available. This cannot be done using wiggle matching techniques.

The statistical framework explained in Chapter 4 allows us to tackle this problem and consistently combine all the factors explained above. Below we present an example of how this can be done.

5.5.1 An example

Clymo *et al.* (1990) present a study on a peat core sampled from Southwest Scotland. The peat core is 50cm deep and samples were taken to be radiocarbon dated approximately every 2cm below the summer water table. Details of these determinations can be found in Table 5.5. Here we present an analysis of these determinations and obtain a description for peat depth and age.

Dry mass is easily measured. For the peat core under study there is a clear linear relationship between cumulative dry mass and depth below the summer water table (see Clymo *et al.* 1990, figure 1). This means that Equation 5.2 can be rewritten as

$$x = x_0 + \frac{\rho'}{\alpha} \{1 - \exp(-\alpha(\theta - \theta_0))\}. \quad (5.3)$$

Here x_0 (= 24cm) is the depth for the summer water table and θ_0 is the age for the peat at that depth. $\rho' = a\rho$ for some fixed a , and ρ and α are defined as in

| j | Depth x_j cm | Sample id. | Det. |
|-----|-------------------|------------|--------------|
| 1 | 26 | UB-2773 | 120 ± 60 |
| 2 | 28 | UB-2774 | 220 ± 60 |
| 3 | 30 | UB-2775 | 110 ± 50 |
| 4 | 32 | UB-2776 | 50 ± 66 |
| 5 | 34 | UB-2777 | 125 ± 60 |
| 6 | 36 | UB-2778 | 205 ± 55 |
| 7 | 38 | UB-2779 | 285 ± 55 |
| 8 | 40 | UB-2780 | 255 ± 55 |
| 9 | 42 | UB-2781 | 110 ± 60 |
| 10 | 44 | UB-3165 | 434 ± 35 |
| 11 | 46 | UB-3166 | 404 ± 31 |

Table 5.5
Radiocarbon determinations from Ellergower Moss core EK3.
Baillie (personal communication).

Equation 5.2.

Therefore, once knowing θ_0 , α and ρ' we would know the relationship between depth and age (below x_0). Unfortunately none of these parameters are known exactly and thus need to be estimated using radiocarbon. Using our framework, it is not difficult to tackle this problem. Here we define

$$\psi = (\theta_0, \alpha, \rho').$$

Then we note, expressing Equation 5.3 in terms of the depths x_j 's, that given ψ ,

$$\theta_j = A_\psi(x_j) = \theta_0 - \frac{1}{\alpha} \log \left\{ 1 - \frac{\alpha}{\rho'} (x_j - x_0) \right\}.$$

Thus, we know *a priori* that $(\theta \mid \psi) = (A_\psi(x_1), A_\psi(x_2), \dots, A_\psi(x_m))$ with probability one. That is, given ψ the θ_j 's are known (since x_0 and x_1, x_2, \dots, x_m are known ie. the depth of the summer water table and the depth for the samples radiocarbon dated).

We use the distribution $f(y | \theta)$ as in Equation 4.1, that is

$$y_j | \theta_j \sim N(\mu(\theta_j), \omega_j^2(\theta_j))$$

(where $\omega_j^2(\theta_j) = \sigma_j^2 + \sigma^2(\theta_j)$) and we assume that, given the θ_j 's, the y_j 's are independent and therefore

$$f(y | \theta) = \prod_{j=1}^m f(y_j | \theta_j).$$

To this point we have defined ψ , $f(y | \theta)$ and $f(\theta | \psi)$ and thus $f(\psi)$ (the prior distribution for ψ) is left to be defined. We expect that the age of peat at the summer water table is independent of the rate of mass accumulation and decay and thus we assume that $f(\psi) = f(\theta_0)f(\alpha, \rho')$. Furthermore, $1 - \frac{\alpha}{\rho'}(x_m - x_0)$ must be greater than zero (so $A_\psi(x)$ is well defined) and since α is a proportion, $f(\alpha, \rho')$ must have support contained in

$$1 > \alpha > 0 \text{ and } \frac{\rho'}{\alpha} > (x_m - x_0).$$

The above parameters have been estimated before for other peat cores, using both formal and informal techniques (see Clymo 1991, Warner *et. al* 1993, Smith and Clymo 1984). In principle, reliable *a priori* information about θ_0 , α and ρ' can be used in the analysis and modelled with the distribution $f(\psi)$. These parameters change for different geographical areas and therefore it is difficult to have information about θ_0 , α and ρ' in a truly general sense. Little can be said about θ_0 , however, from previous experience and in very broad terms, typical values for α and ρ (in the northern hemisphere) are $\alpha = 1 \times 10^{-4}$ to $5 \times 10^{-4} \text{ yr}^{-1}$ and $\rho = 0.003$ to $0.007 \text{ gr cm}^{-2} \text{ yr}^{-1}$. It will be very difficult to decide whether these figures apply to our example or not. With this in mind, and considering that $\rho' = a\rho$ where $a \approx 10 \text{ gr}^{-1} \text{ cm}^3$, we have decided to give α and ρ' a uniform prior distribution over the (very) wide margins $0.01 > \alpha > 0$ and $0.3 > \rho' > 0$

(with $\frac{\rho'}{\alpha} > (x_m - x_0)$), and we let $f(\theta_0)$ be vague.

From the above discussion it is not difficult to see that the posterior distribution for ψ is

$$f(\psi | y) = f(\theta_0, \alpha, \rho' | y) \propto \prod_{j=1}^m \frac{1}{\omega_j(A_\psi(x_j))} \exp \left\{ -\frac{1}{2} \frac{(\mu(A_\psi(x_j)) - y_j)^2}{\omega_j(A_\psi(x_j))} \right\}$$

for $0.01 > \alpha > 0$, $0.3 > \rho' > 0$ and $\frac{\rho'}{\alpha} > (x_m - x_0)$. We used numerical integration techniques to find $f(\psi | y)$ and the posterior marginals $f(\theta_0 | y)$ and $f(\alpha, \rho' | y)$. We present plots of these posterior marginals in Figures 5.10 and 5.11 respectively. The 95% HPD region for θ_0 , the age for the peat at the summer water table, is 250 to 85 BP and the mode for the distribution (the MAP estimator) is at 212 BP.

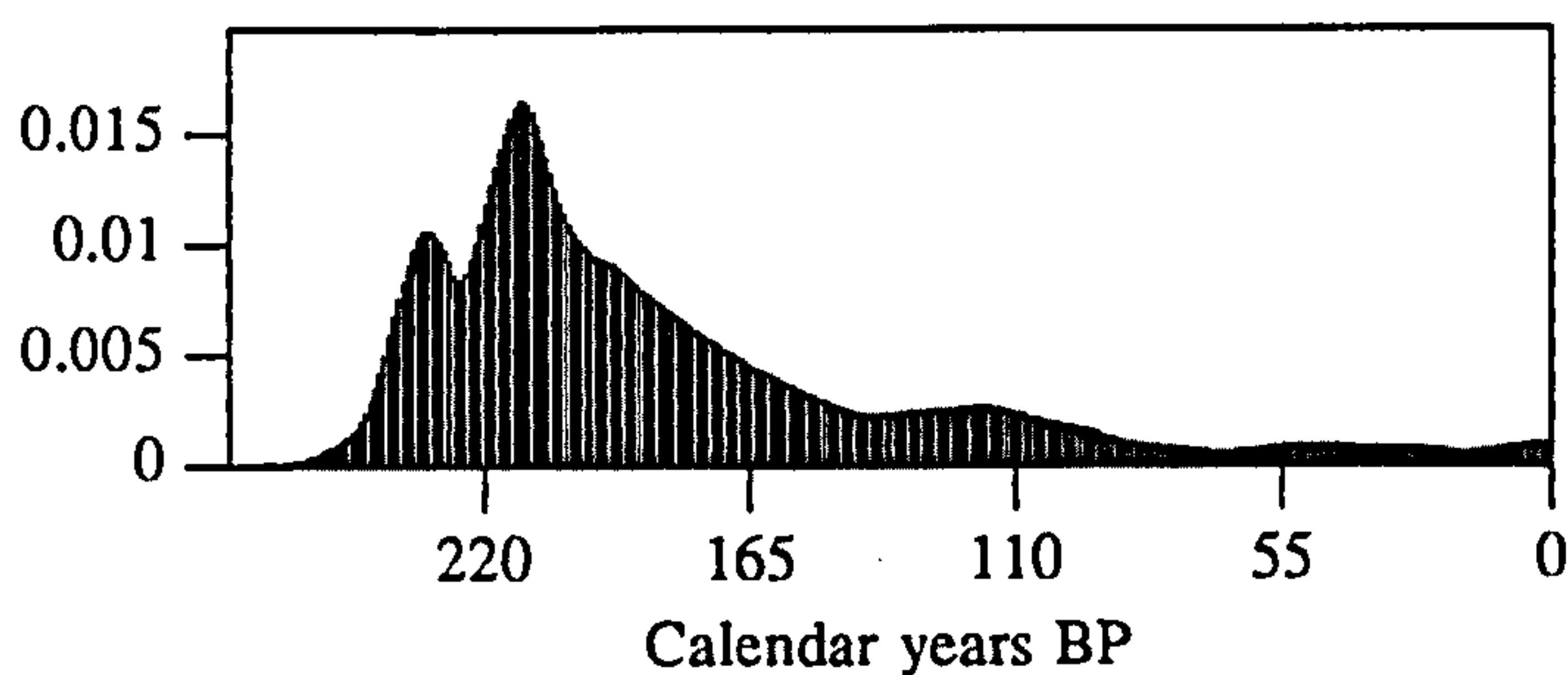


Figure 5.10
Posterior distribution for the age of the peat
at the summer water table.

Note that the joint posterior distribution for α and ρ' is wide spread over the area chosen for the prior distribution of those parameters. However, there is a high peak in the region $0 < \alpha < 0.004$ and $0.05 < \rho' < 0.1$. The mode for the whole distribution is (0.0016, 0.76). This translates to $\alpha = 0.0016$ and $\rho = 0.076$. The point estimate for ρ approximately coincides with previous values reported for that parameter. However, our point estimate for α is 2 to 3 times higher than values previously reported. Experts on the subject might argue that a part of the

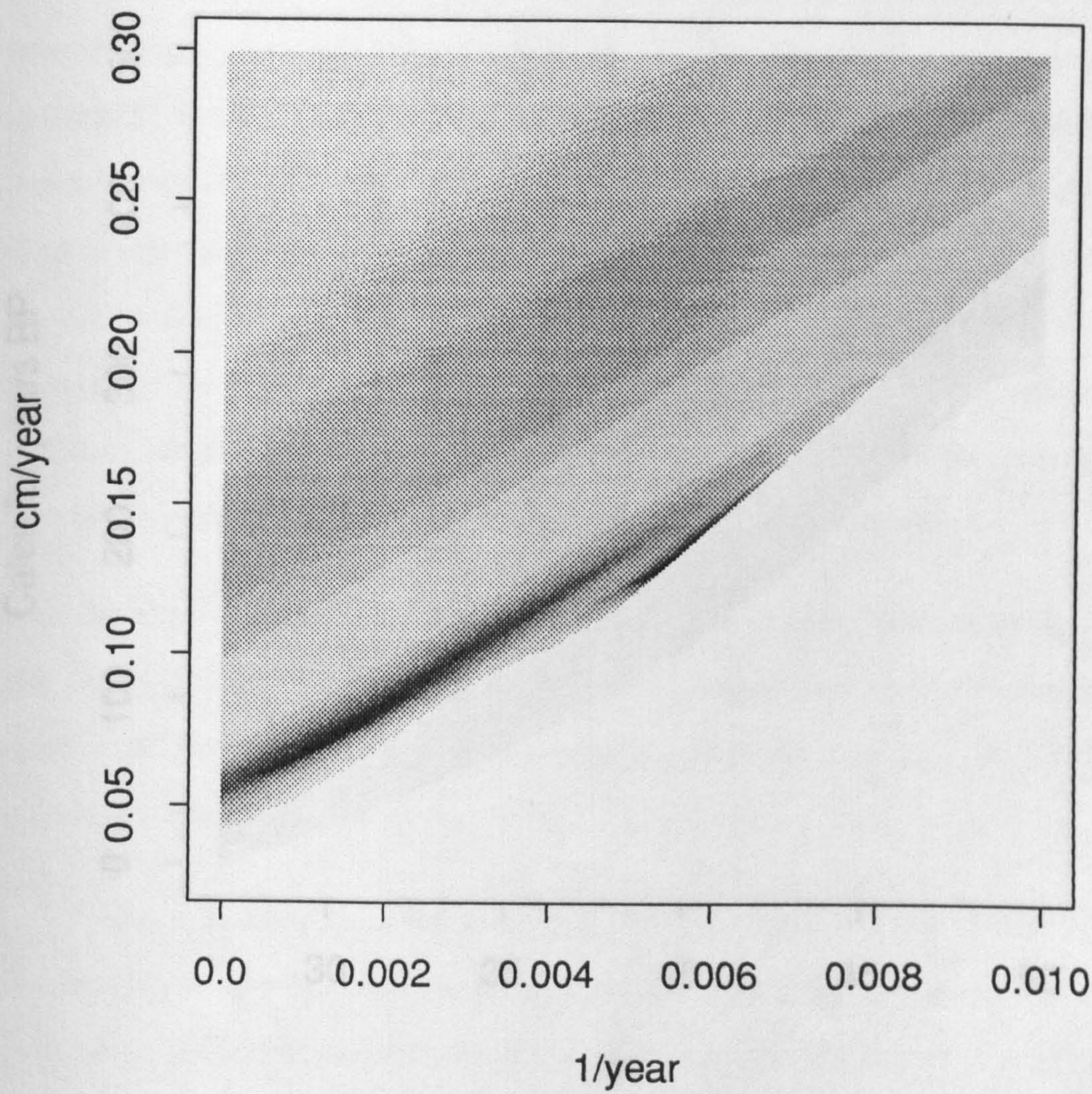


Figure 5.11
Joint posterior distribution for α (x-axis) and ρ' (y-axis).

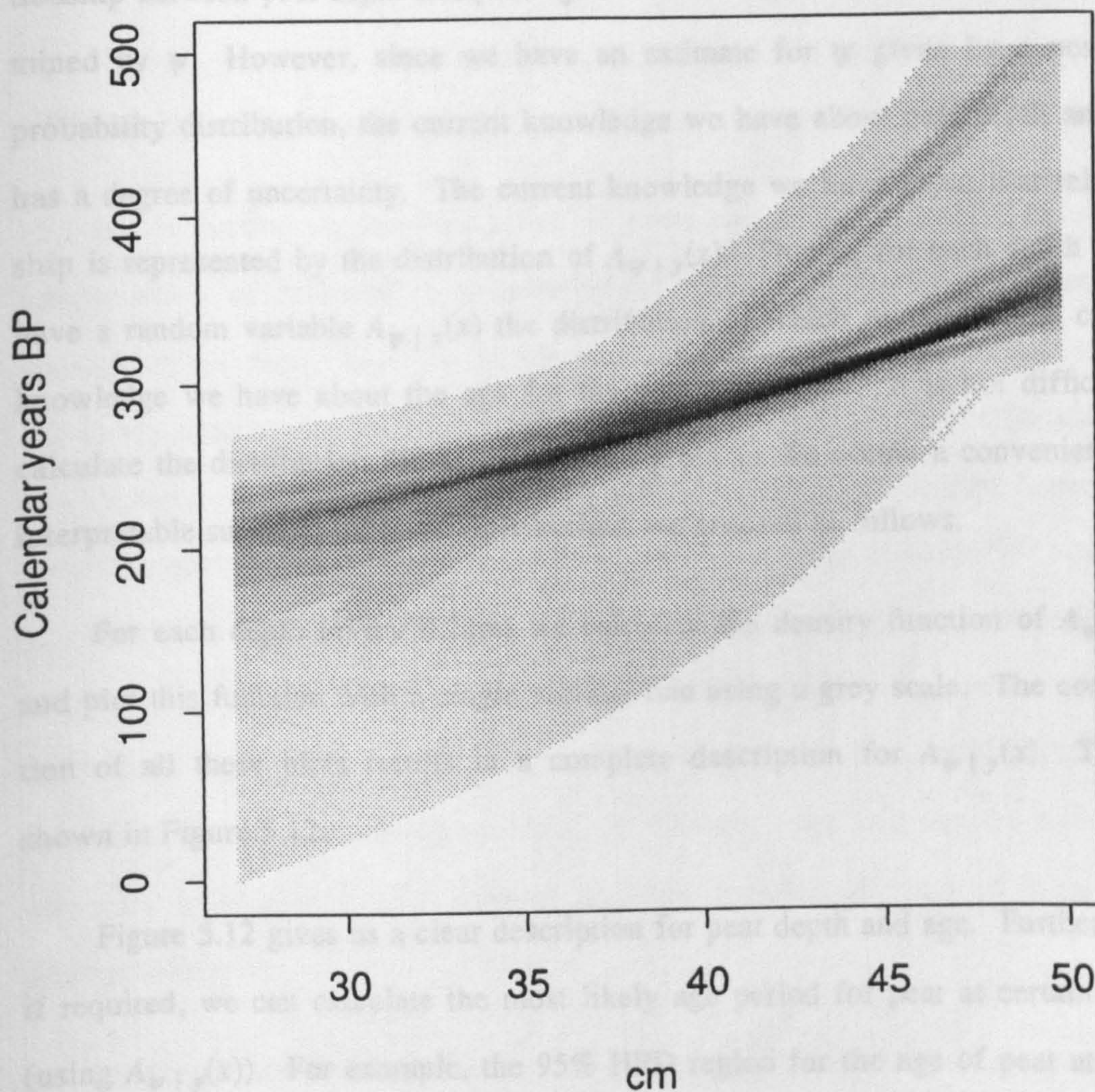


Figure 5.12
 Plot describing the relationship between peat depth and age.
 A vertical line at depth x describes, based on the gray scale,
 the posterior density function for the age of the peat at that depth.

a priori region chosen for α and ρ' represents physically impossible values and thus need to be reduced to a more sensible region. Further collaboration with the experts involved should result in more satisfactory prior elicitation.

As stated before, the principal objective of this study is to establish the relationship between peat depth and peat age. In our model this relationship is determined by ψ . However, since we have an estimate for ψ given by a posterior probability distribution, the current knowledge we have about peat depth and age has a degree of uncertainty. The current knowledge we have about that relationship is represented by the distribution of $A_{\psi | y}(x)$. That is, for each depth x , we have a random variable $A_{\psi | y}(x)$ the distribution of which represents the current knowledge we have about the age for the peat at depth x . It is not difficult to calculate the distribution for $A_{\psi | y}(x)$ using $f(\psi | y)$. To obtain a convenient and interpretable summary of these distributions we proceed as follows.

For each depth (every 0.2cm) we calculate the density function of $A_{\psi | y}(x)$ and plot this function with a *single* vertical line using a grey scale. The conjunction of all these plots results in a complete description for $A_{\psi | y}(x)$. This is shown in Figure 5.12.

Figure 5.12 gives us a clear description for peat depth and age. Furthermore, if required, we can calculate the most likely age period for peat at certain depth (using $A_{\psi | y}(x)$). For example, the 95% HPD region for the age of peat at depth 35cm is approximately 300 to 200 years BP.

We believe that the above approach for analysing the age of peat cores can give the environmental scientists reliable and interpretable results. Note that, using our framework, the important factors involved in this problem have been introduced into the analysis. Namely, radiocarbon calibration (included in the model $f(y | \theta)$), the theoretical considerations about peat growth and the

radiocarbon determinations available. Further research could analyse other data sets available from various parts of the world. In addition, other models for peat growth ($A_{\psi}(x)$ above) could be considered and, indeed, the priors used (eg. $f(\alpha, \rho')$) can be changed to be of use in other peat core studies. We are confident that this will help in the understanding of peat growth.

5.6 Working with phases

In this Section we present a series of techniques to tackle the problems of the interpretation of radiocarbon determinations when complicated 'archaeological phases' are involved. Here we do not base our study on a particular archaeological dating problem but try to consider general techniques that could be used. In Naylor and Smith (1988) and Buck *et al.* (1992), reviewed in Chapter 2, examples involving complicated archaeological phase structures have been studied and complex models have been used to date them by combining the prior information and the available radiocarbon determinations. In this Section we present a series of techniques to deal with phase structures such as those that could appear in a specific archaeological dating problem. In Chapter 7 we will use the techniques presented here to aid in the analysis and interpretation of a set of radiocarbon determinations.

In broad terms, the archaeological dating problem studied is the following. There are a series of time periods, or phases, identified by the archaeologists for the problem under study, that need to be dated using radiocarbon. These could be a series of stratified periods of occupation in a site, a series of periods of stylistically similar pottery production, the reign of a series of kings in a given culture, etc. For each phase there is a set of radiocarbon determinations associated with it. Furthermore, prior knowledge might be available about relative chronological orders among the phases and/or about their lengths or absolute position in time.

It is then expected that, by combining all of these factors, the phases could be dated in the calendar scale. Below we give further details and explain how this problem can be tackled using our framework.

5.6.1 A general setting

Here we present a general setting for the analysis of complicated phase structures. Firstly, we give some notational conventions and explain the assumptions made. Secondly, we present an example and go on to discuss the general sampling techniques needed to be able to approximate the posterior distributions of interest.

Suppose that we have a set of m radiocarbon determinations and that there are n phases which need to be dated. Let $y_{i,j} \pm \sigma_{i,j}$ be determination j associated with phase i , and m_i be the number of determinations associated with phase i (thus $\sum_{i=1}^n m_i = m$). Let $\theta_{i,j}$ be the associated calendar year for determination $y_{i,j}$. We define the vector

$$\theta_i = (\theta_{i,1}, \theta_{i,2}, \dots, \theta_{i,m_i}),$$

with an equivalent definition for y_i . Furthermore, we define θ as the vector containing all the $\theta_{i,j}$'s and y the vector containing all the $y_{i,j}$'s. We use the term 'phase' to mean a 'period of time'. In terms of our framework, we will represent the archaeological phases which need to be dated using the parameter ψ .

Now, let ψ_{2i} be the beginning and ψ_{2i-1} the end of phase i , for $i = 1, 2, \dots, n$. We assume that, for a fixed i , the associated calendar years $\theta_{i,j}$'s all belong to the phase ψ_{2i} to ψ_{2i-1} . That is, the calendar year in which the organic matter in sample i,j died belongs to the phase which the sample is related to (phase i).

Here, considering our general framework, we define

$$\psi = (\psi_1, \psi_2, \dots, \psi_{2n}).$$

That is, ψ represents the periods of time (in the calendar scale) covered by the phases under study.

It is common with problems of this kind to assume that the beginning or end of one phase coincides with the beginning or end of another phase or phases. This means that, in some examples, some ψ_i 's will be the same. To represent this we use the set $e(i)$, where $k \in e(i)$ if and only if $\psi_i = \psi_k$.

The problem now is to define $f(y | \theta)$, $f(\theta | \psi)$ and $f(\psi)$ and, given these, obtain the posterior distribution $f(\psi | y)$. This posterior distribution will represent the knowledge we have about the position in time for the phases under study, considering our prior information (given by $f(\theta | \psi)$ and $f(\psi)$) and the radiocarbon determinations y . Thus the primary objective of our analysis will be to calculate $f(\psi | y)$.

Using the model in Equation 4.1 we have

$$y_{i,j} | \theta_{i,j} \sim N(\mu(\theta_{i,j}), \omega_{i,j}^2(\theta_{i,j}))$$

(where $\omega_{i,j}^2(\theta_{i,j}) = \sigma_{i,j}^2 + \sigma^2(\theta_{i,j})$). We assume that, given the $\theta_{i,j}$'s, the $y_{i,j}$'s are independent and thus

$$f(y | \theta) = \prod_{i=1}^n \prod_{j=1}^{m_i} f(y_{i,j} | \theta_{i,j}).$$

From the above, it is clear that, given ψ_{2i} and ψ_{2i-1} , $\theta_{i,j} \in (\psi_{2i-1}, \psi_{2i})$ with probability one. From this it seems reasonable to assume that $f(\theta_{i,j} | \theta_{-i,j}, \psi) = f(\theta_{i,j} | \theta_{-i,j}, \psi_{2i-1}, \psi_{2i})$. In addition, there might be cases in which the associated calendar years $\theta_{i,j}$ are known, *a priori*, to maintain some chronological

order. This information can be introduced in our analysis when we specify $f(\theta_{i,j} \mid \theta_{-i,j}, \psi_{2i-1}, \psi_{2i})$ (for example, given $\theta_{-i,j}, \theta_{i,j+1} > \theta_{i,j} > \theta_{i,j-1}$). However, here we assume that

$$f(\theta_{i,j} \mid \theta_{-i,j}, \psi_{2i-1}, \psi_{2i}) = f(\theta_{i,j} \mid \psi_{2i-1}, \psi_{2i}).$$

Furthermore, it seems reasonable to expect that, knowing the position in time for the phases, the associated calendar years are independent. That is, we assume that given ψ , the $\theta_{i,j}$'s are independent. This means that

$$f(\theta \mid \psi) = \prod_{i=1}^n \prod_{j=1}^{m_i} f(\theta_{i,j} \mid \psi_{2i-1}, \psi_{2i}). \quad (5.4)$$

The actual definition for $f(\theta_{i,j} \mid \psi_{2i-1}, \psi_{2i})$ will obviously depend on the problem at hand. If we have little or no information that indicates that the organic matter contained in sample i,j should have died in a specific part of the phase i , then we can assume that

$$\theta_{i,j} \mid \psi \sim U(\psi_{2i-1}, \psi_{2i}).$$

That is, sample i,j is, *a priori*, equally likely to have died at any time within phase i . If, for example, sample i,j is thought to have died near to the end of phase i , then $f(\theta_{i,j} \mid \psi_{2i-1}, \psi_{2i})$ can be chosen as a skewed distribution towards ψ_{2i-1} . However, we will assume a uniform prior for $\theta_{i,j}$. With this in mind, in Figure 5.13 we present the hierarchy diagram for the parameters involved in phase i , namely, $\theta_{i,j}$ for $j = 1, 2, \dots, m_i$ and ψ_{2i-1} and ψ_{2i} .

As discussed above, there may be further chronological prior information relating the beginning and end of the phases under study. For example, we may know, *a priori*, that phase i ended before phase $i+1$ began. That is, we know that $\psi_{2i-1} > \psi_{2(i+1)}$; see Figure 5.14. This information can be introduced in our

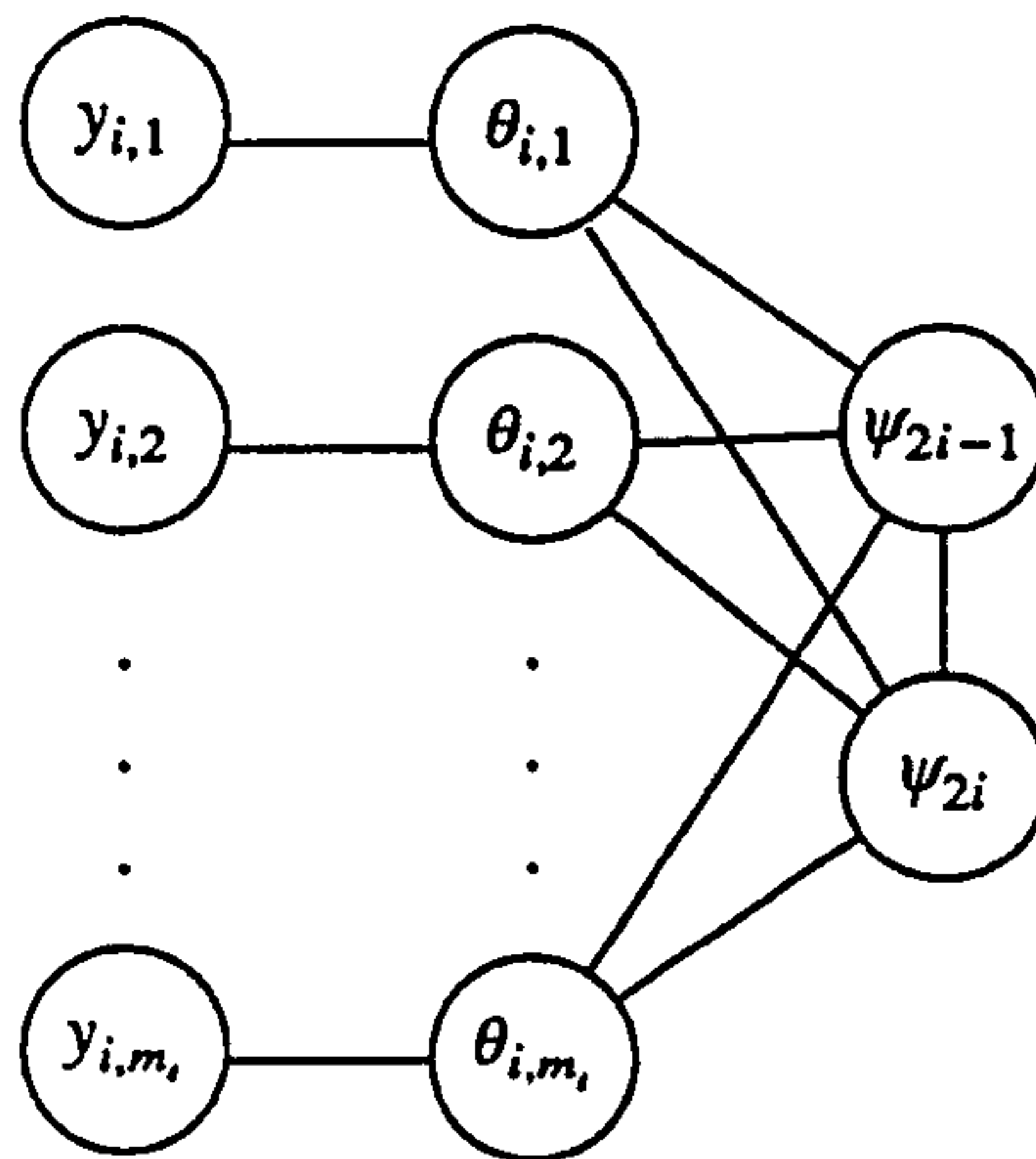


Figure 5.13

Hierarchy diagram for the parameters involved in phase i .

analysis by defining

$$\psi_{2i-1} \mid \psi_{-(2i-1)} \sim U(\psi_{2(i+1)}, \psi_{2i})$$

(note that, since ψ_{2i-1} is the end of phase i , we should always have $\psi_{2i-1} < \psi_{2i}$).

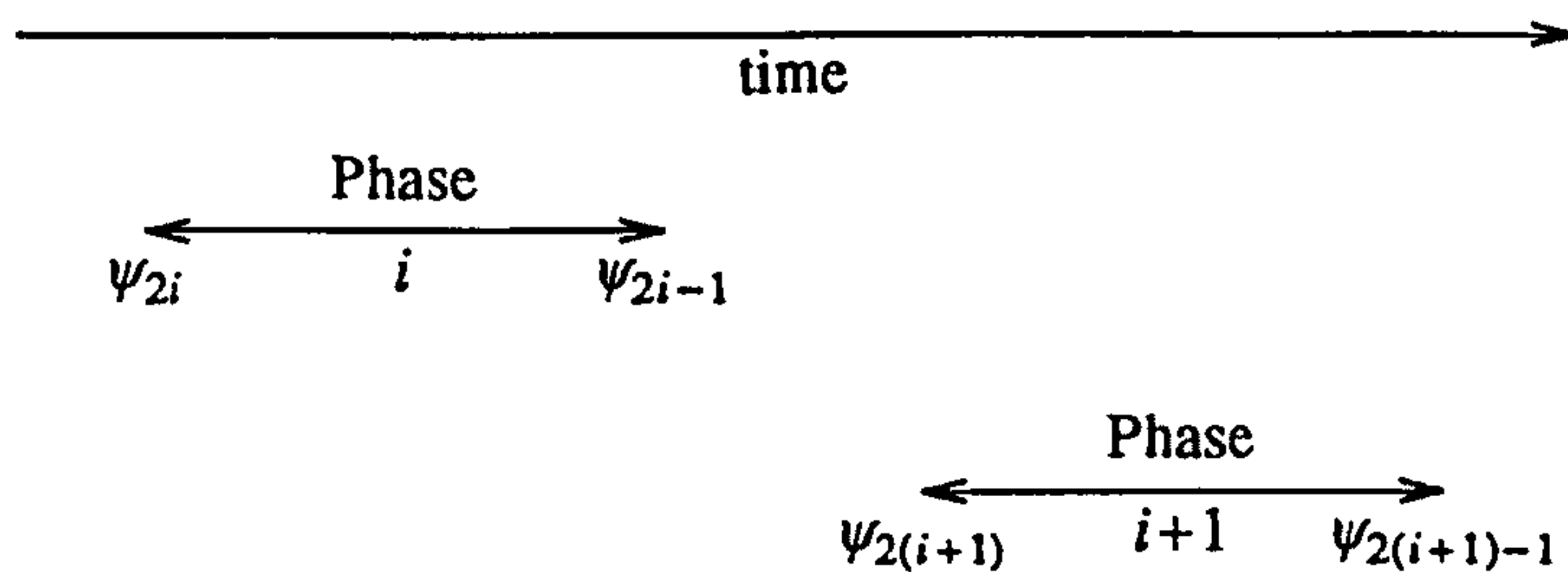


Figure 5.14

Diagram representing the relative position for the beginning and end for two phases. Here the end of phase i is before the beginning of phase $i+1$.

A very general way of representing chronological relationships among the ψ_i 's like the ones discussed above, is to assume that

$$(\psi_i \mid \psi_{-i}) \sim U(a_i(\psi_{-i}), b_i(\psi_{-i})).$$

This means that, *a priori*, ψ_i depends upon ψ_{-i} only through the functions a_i and

b_i , which are the boundaries for the prior uniform distribution of ψ_i . That is, given ψ_{-i} , ψ_i is uniformly distributed within the period of time $b_i(\psi_{-i})$ to $a_i(\psi_{-i})$.

The above assumptions may restrict the wide range of possibilities of the general framework stated in Chapter 4. However, from our experience of some archaeological dating problems studied we see that the above setting gives a wide enough range of possibilities. In fact, the archaeological phases problems studied by Naylor and Smith (1988) and Buck *et al.* (1992) (and the phases problem studied in Chapter 7) can be included within the above setting.

5.6.2 Form of the full conditionals

Using the general setting discussed above and using Equations 4.2 and 4.3, it is easy to see that the full conditionals for our parameters are,

$$f(\theta_{i,j} \mid y, \theta_{-(i,j)}, \psi) = f(\theta_{i,j} \mid y_{i,j}, \psi_{2i-1}, \psi_{2i}) \propto \frac{1}{\omega_{i,j}(\theta_{i,j})} \exp \left\{ -\frac{1}{2} \left(\frac{(y_{i,j} - \mu(\theta_{i,j}))^2}{\omega_{i,j}^2(\theta_{i,j})} \right) \right\} I(\theta_{i,j})_{(\psi_{2i-1}, \psi_{2i})} \quad (5.5)$$

where $\omega_{i,j}^2(\theta_{i,j}) = \sigma_{i,j}^2 + \sigma^2(\theta_{i,j})$, and

$$f(\psi_i \mid y, \theta, \psi_{-i}) = f(\psi_i \mid \theta, \psi_{-i}) \propto f(\theta \mid \psi) I(\psi_i)_{(a_i(\psi_{-i}), b_i(\psi_{-i}))}.$$

Note that the full conditionals $f(\theta_{i,j} \mid y_{i,j}, \psi_{2i-1}, \psi_{2i})$ are similar to the distributions for the calibrated calendar year of a determination $y \pm \sigma$, with vague prior, found in Chapter 3 (see Figures 3.5 or 3.6), but restricted to (ψ_{2i-1}, ψ_{2i}) .

To analyse the form of the full conditionals for the ψ_i 's we look at the product $f(\theta \mid \psi) I(\psi_i)_{(a_i(\psi_{-i}), b_i(\psi_{-i}))}$. To facilitate notation we call $g(\psi_i) =$

$f(\psi_i | \theta, \psi_{-i})$. From Equation 5.4 we see that $f(\theta | \psi)$ is a product of uniform densities. From those uniform densities, we only consider the ones that include ψ_i as a parameter. Now note that, if $i = 2k-1$ then

$$\prod_{j=1}^{m_k} f(\theta_{i,j} | \psi_i, \psi_{2k}) \propto (\psi_{2k} - \psi_i)^{-m_k} \prod_{j=1}^{m_k} I(\theta_{i,j})_{(\psi_i, \psi_{2k})}$$

with a similar expression if $i = 2k$. We must remember that ψ_i can represent the beginning and end of several phases. The set $e(i)$ tells us that $2k \in e(i)$ if and only if $\psi_i = \psi_{2k}$ and thus ψ_i is the beginning of phase k and that $2k-1 \in e(i)$ if and only if $\psi_i = \psi_{2k-1}$ and, therefore, ψ_i is the end of phase k . From this it is not difficult to see that

$$\begin{aligned} g(\psi_i) = K \times \\ \prod_{2k \in e(i)} \left\{ (\psi_i - \psi_{2k-1})^{-m_k} \prod_{j=1}^{m_k} I(\theta_{k,j})_{(\psi_{2k-1}, \psi_i)} \right\} \times \\ \prod_{(2k-1) \in e(i)} \left\{ (\psi_{2k} - \psi_i)^{-m_k} \prod_{j=1}^{m_k} I(\theta_{k,j})_{(\psi_i, \psi_{2k})} \right\} \times \\ I(\psi_i)_{(a_i(\psi_{-i}), b_i(\psi_{-i}))}. \end{aligned}$$

Here the first set of products accounts for when ψ_i is the end of a phase and the second set for when ψ_i is the beginning of a phase, and K is the normalisation constant. Now we need to express the above indicator functions $I(\theta_{k,j})_{(\psi_{2k-1}, \psi_i)}$, $I(\theta_{k,j})_{(\psi_i, \psi_{2k})}$ and $I(\psi_i)_{(a_i(\psi_{-i}), b_i(\psi_{-i}))}$ in terms of ψ_i , after which we obtain

$$g(\psi_i) = K \left\{ \prod_{2k \in e(i)} (\psi_i - \psi_{2k-1})^{-m_k} \prod_{(2k-1) \in e(i)} (\psi_{2k} - \psi_i)^{-m_k} \right\} I(\psi_i)_{(a, b)},$$

where

$$a = \max \left\{ \max_{2k \in e(i)} (\theta_k), a_i(\psi_{-i}) \right\},$$

$$b = \min \left\{ \min_{(2k-1) \in e(i)} (\theta_k), b_i(\psi_{-i}) \right\}.$$

Therefore, we can see that $g(\psi_i)$ has the form

$$g(\psi_i) = K \left\{ s \prod_l (\psi_l - \psi_i)^{-q_l} \right\} I(\psi_i)_{(a, b)}, \quad (5.6)$$

where $s = 1$ or -1 (arising from the changes of sign due to switching the $(\psi_i - \psi_{2l})$'s to $(\psi_{2l} - \psi_i)$), q_l are some positive integers and K is the normalisation constant.

From the above considerations we can conclude that

$$f(\psi_i | y, \theta, \psi_{-i}) = \frac{K}{P(\psi_i)} I(\psi_i)_{(a, b)}$$

where $P(\psi_i)$ is a polynomial positive on $[a, b]$, to ensure integration, with its roots at some ψ_k 's.

Note that for some cases, there might be parameters that are actually 'repeated' in our setting. If for example, the beginning of phase i is the ending of phase $i+1$ then we have $\psi_{2i} = \psi_{2(i+1)-1}$. We prefer, though, to maintain both parameters and make our modelling more explicit. We call the complete vector ψ the 'explicit' parameters, and the corresponding vector ψ' , for which all the parameters are different, the 'actual' parameters.

5.6.3 An example

We present an example to illustrate the general setting discussed above. We use a hypothetical phases problem to illustrate our approach. Suppose we have four archaeological phases in a site and assume that the prior distribution available can be represented with the priors explained above. Suppose, for example, that the beginning and end for each phase are known, *a priori*, to maintain a

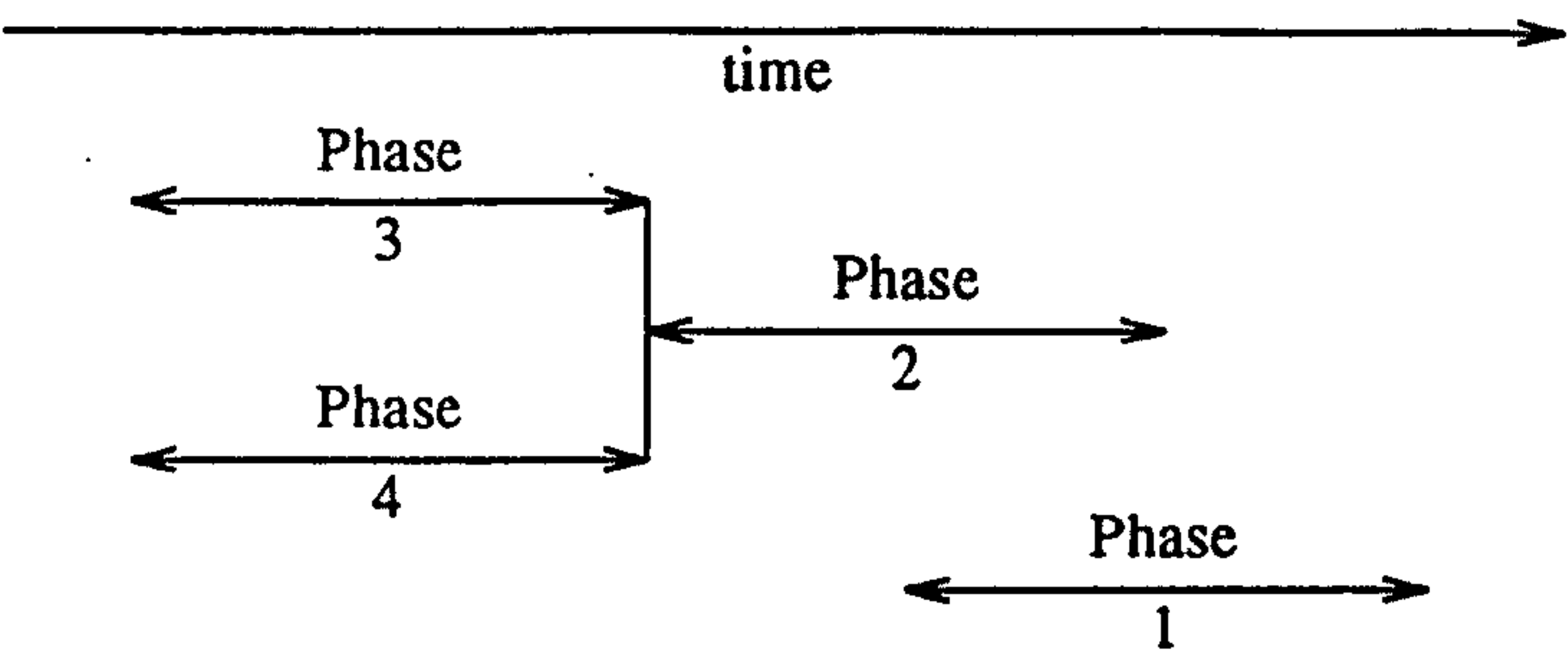


Figure 5.15

Diagram representing the known, *a priori*, relative chronological relationships among the phases in our hypothetical example. The vertical line means that the end of phases 3 and 4 coincides with the beginning of phase 2.

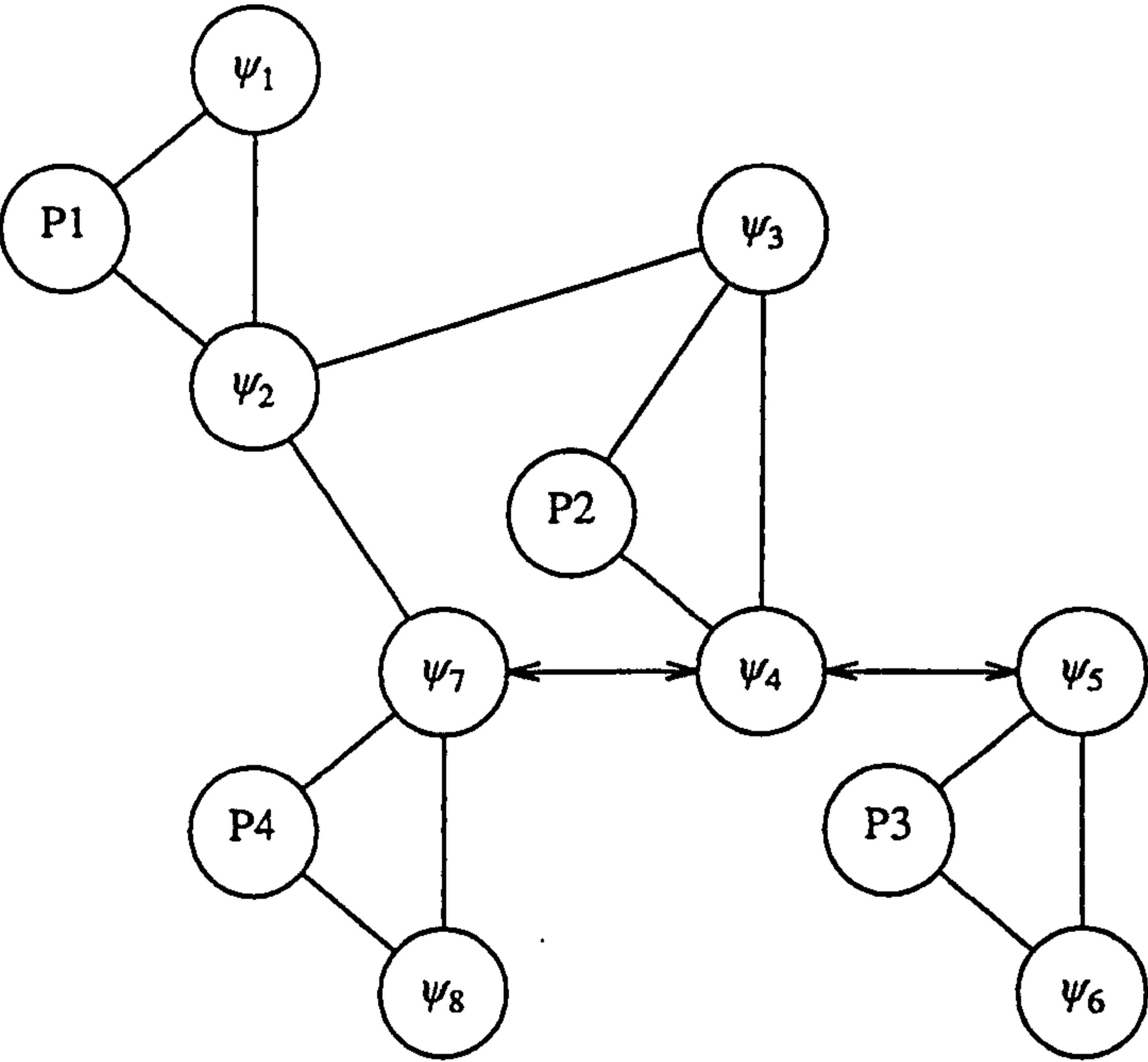


Figure 5.16

Hierarchy diagram for our hypothetical example.

The nodes $P1$, $P2$, $P3$ and $P4$ represent the $y_{i,j}$'s and the $\theta_{i,j}$'s in phases $P1$, $P2$, $P3$ and $P4$ respectively.

relative chronological order, which we represent by the diagram in Figure 5.15.

Using the notation presented above, the hierarchy diagram for the parameters in this example can be seen in Figure 5.16. To abbreviate our diagram we use the nodes $P1$, $P2$, $P3$ and $P4$ which represent the $y_{i,j}$'s and the $\theta_{i,j}$'s in phases

$P1, P2, P3$ and $P4$ respectively. (We have described the dependencies among these latter parameters in Figure 5.14.)

Here time runs upwards in our hierarchy diagram, thus, for example, in the diagram presented in Figure 5.16, ψ_7 is earlier than ψ_2 ($\psi_7 > \psi_2$) etc. This will be true, though, only in between nodes sharing a link. Double arrowed linkages mean that the parameters are equal, thus $\psi_4 = \psi_5 = \psi_7$ and vertical links are only left to link the boundaries of groups. Therefore we have $\psi_3 < \psi_2$, $\psi_2 < \psi_7$ and $\psi_4 = \psi_5 = \psi_7$. These chronological relationships can be introduced in our analysis using the prior distributions $f(\psi_i | \psi_{-i})$ defined previously. For example

$$\psi_2 | \psi_{-2} \sim U(\max(\psi_3, \psi_1), \psi_7).$$

Here, of course, $a_2(\psi_{-2}) = \max(\psi_3, \psi_1)$ and $b_2(\psi_{-2}) = \psi_7$.

We choose our 'actual' set of parameters to be

$$\psi' = (\psi_1, \psi_2, \psi_3, \psi_4, \psi_6, \psi_8).$$

After this and following the above discussion on the form of the full conditionals, it is not difficult to find the full conditionals for our parameters. For example, the full conditional for ψ_4 depends on $\max(\theta_2)$, ψ_3 and, since $\psi_4 = \psi_5 = \psi_7$, it also depends on $\min(\theta_3)$, $\min(\theta_4)$, ψ_2 , ψ_6 and ψ_8 (see Figure 5.16). From the conclusions of the previous Section we can see that

$$f(\psi_4 | \theta, \psi_{-4}) \propto (\psi_4 - \psi_3)^{-m_2} (\psi_6 - \psi_4)^{-m_3} (\psi_8 - \psi_4)^{-m_4} I(\psi_4)_{(a, b)}$$

where $a = \max(\max(\theta_2), \psi_2)$ and $b = \min(\theta_3, \theta_5)$. The rest of the full conditionals can be found in a similar way.

5.6.4 Sampling techniques

Considering the above discussion on the form of the full conditionals, we only need to sample from two types of distributions whose densities are given in Equations 5.5 and 5.6. These are the full conditionals for the $\theta_{i,j}$'s and the full conditionals for the ψ_i 's.

In previous Sections we have discussed how to sample from the distributions $f(\theta_{i,j} \mid y, \theta_{-i,j}, \psi)$ (see Section 4.3.5). In this case we need, either to integrate numerically to find the cumulative distribution function $F(\theta_{i,j})$ and then find $\theta_{i,j}^{(*)}$ such that $G(\theta_{i,j}^{(*)}) = u$, where $U \sim U(0, 1)$, or to use a rejection type algorithm for the simulation. As we have pointed out before, analytical methods are discarded due to the non-analytical definition of the calibration curve $\mu(\theta)$ and the variance function $\sigma^2(\theta)$. However, we do not discuss this problem any further.

We then concentrate our attention on sampling from the distributions $f(\psi_i \mid y, \theta, \psi_{-i})$. If we let $g(\psi_i) = f(\psi_i \mid y, \theta, \psi_{-i})$, we have seen that

$$g(\psi_i) = \frac{K}{P(\psi_i)} I(\psi)_{(a, b)} \quad (5.7)$$

where $P(\psi_i)$ is a polynomial with the form $s \prod_l (\psi_l - \psi_i)^{q_l}$ positive in $[a, b]$ with $s = 1$ or -1 and q_l some positive integers (see Equation 5.6). Elsewhere we have seen how to sample from $g(\psi_i)$ when $P(\psi_i) = (\psi_{2k} - \psi_i)^{m_k}$ (see Section 4.3.5). In this case it is possible to find the inverse of the cumulative distribution function of ψ_i and from that we have

$$\psi_i^{(*)} = \psi_{2k} - \frac{\psi_{2k} - b}{\left\{ \left(\frac{\psi_{2k} - b}{\psi_{2k} - a} \right)^{m_i - 1} (1 - u) + u \right\}^{\frac{1}{m_i - 1}}}$$

where $U \sim U(0, 1)$. Then by simulating a value u from a uniform distribution $U(0, 1)$, we obtain a simulated value $\psi_i^{(*)}$ for ψ_i . Similarly, when $P(\psi_i) =$

$(\psi_i - \psi_{2k-1})^{m_k}$ we have

$$\psi_i^{(*)} = \psi_{2k-1} - \frac{a - \psi_{2k-1}}{\left\{ \left(\frac{a - \psi_{2k-1}}{b - \psi_{2k-1}} \right)^{m-1} (1-u) + u \right\}^{\frac{1}{m_i-1}}}$$

where $U \sim U(0, 1)$.

For the general case $P(\psi_i) = s \prod_l (\psi_l - \psi_i)^{q_l}$, it is not possible to find an analytical expression for the inverse of the cumulative distribution function and thus to obtain similar results to the ones above. Nevertheless, it is very simple (though it might not be very efficient) to perform a rejection type algorithm to simulate a value for ψ_i , given the following Lemma.

Lemma: The maximum of $g(\psi_i)$ in Equation 5.7 is either at a or at b .

Proof: for the purposes of this proof and with no loss of generality, we let

$$P(\psi_i) = \prod_{k=1}^n (\psi_k - \psi_i)^{q_k}.$$

Since $g(\psi_i)$ is a density function, $P(\psi_i) > 0$ in $[a, b]$. We want to prove that the minimum of $P(\psi_i)$ in $[a, b]$ is either at a or at b since $g(\psi_i) = \frac{K}{P(\psi_i)} I(\psi)_{(a, b)}$.

Given the above definition for $P(\psi_i)$, it is clear that the ψ_k 's are the roots of $P(\psi_i)$, each with multiplicity q_k . With no loss of generality we can assume that $\psi_1 < \psi_2 < \dots < \psi_n$. From this it is clear that the derivative of $P(\psi_i)$, $P'(\psi_i)$, has a root in (ψ_k, ψ_{k+1}) for $k = 1, 2, \dots, n-1$ since $P(\psi_k) = 0$ and $P(\psi_{k+1}) = 0$. Each of the ψ_k 's is a root of $P'(\psi_i)$ of multiplicity $q_k - 1$, this means that $P'(\psi_i)$ has at least

$$n-1 + \sum_{k=1}^n (q_k - 1) = \left(\sum_{k=1}^n q_k \right) - 1$$

real roots. But these are *all* the roots of $P'(\psi_i)$ and therefore $P'(\psi) = 0$ has one

and only one root in (ψ_i, ψ_{i+1}) for $k = 1, 2, \dots, n-1$.

Note that, if $b < \psi_1$ or $\psi_n < a$, then $[a, b]$ is not contained in an interval (ψ_k, ψ_{k+1}) and thus $P'(\psi_i)$ has no roots in $[a, b]$. Therefore, the minimum of $P(\psi_i)$ is either at a or at b . Suppose then that $[a, b]$ is contained in one of the intervals (ψ_k, ψ_{k+1}) ; let this interval be (ψ_l, ψ_{l+1}) .

Now, suppose that $P(\psi_i)$ is minimal at x in (a, b) . Therefore x is the only root of $P'(\psi_i)$ in (ψ_l, ψ_{l+1}) . Note that (since $P(\psi_l) = 0$ and $P(\psi_i) > 0$ in (a, b)) $P(a) > P(x) > P(\psi_l)$ with $\psi_l < a < x$. From this we see that there exists a $x' \in (\psi_l, x)$ such that $P(x') = P(x)$ and therefore $P'(\psi_i)$ has a root in (x', x) . But this is not possible since x is the only root for $P'(\psi_i)$ in (ψ_l, ψ_{l+1}) . This means that $P(\psi_i)$ cannot have a minimum in (a, b) and therefore the minimum for $P(\psi_i)$ in $[a, b]$ is either at a or at b .

Using the above result it is very simple to simulate a value for ψ . We just follow the simplest case of the rejection algorithm.

(i) Take $M = \max(g(a)/K, g(b)/K)$.

(ii) Generate $r_1 \sim U(0, 1)$ and $r_2 \sim U(0, 1)$.

(iii) If $Mr_2 < g(a + (b-a)r_1)/K$ take $\psi^{(*)} = a + (b-a)r_1$ as a simulated value for ψ_i , otherwise repeat (ii) and (iii).

The above procedure is efficient because we do not need to normalise $g(\psi_i)$, since we only need the definition for $g(\psi_i)/K$, and because we do not need to find the maximum of $g(\psi_i)$ using a numerical method, which is normally time-consuming. After having a set of suitable software routines and data structures, the implementation of a MCMC method to find the posterior distributions of

interest turns out not to be all that difficult, at least for problems that match with the setting explained here. This will give us more time to concentrate our attention on, among other factors, the modelling process and assessing the convergence of the parameters in the MCMC method used.

Here we have presented a series of techniques to tackle the statistical problems of dating archaeological phases using radiocarbon. A very wide range of possibilities is covered and we have shown how to calculate the full conditional distributions for the parameters involved and perform the sampling. This will give us the posterior distribution $f(\psi | y)$, that represents the current knowledge about the position in time for the phases under study. In the study of specific examples the assumptions made here should be analysed and, if necessary, corrected to be appropriate for the characteristics of the problem. This might result, however, in more complex full conditionals and sampling techniques. We believe that the study of archaeological phase structures using radiocarbon dating techniques will benefit from the ideas presented here.

Chapter 6

Robustness

6.1 Introduction

In this Chapter we study the robustness of our statistical framework to the presence of outliers. Firstly, we will see that radiocarbon dating is especially prone to the generation of outliers and thus it is desirable that the statistical analyses used for the interpretation of radiocarbon determinations must allow for their presence. Secondly, in Section 6.2 we present a general review in which we briefly discuss the different statistical approaches used to detect outliers. Among these we will characterise the outlier problem in radiocarbon dating and identify an approach that is suitable for radiocarbon dating.

In Section 6.3 we then present a novel approach to the problem of outliers in radiocarbon dating. It is based on a new ‘slippage’ model that extends our basic model for radiocarbon determinations and is intended to help in the identification of outliers. We then use this new model to develop our ‘extended framework’ that now allows for outliers. Although the fundamental idea used to extend our model has been widely explored in the statistics literature, here we propose an original generalisation of the ‘location-shift’ modelling approach for the identification of outliers.

In this Chapter we present two examples to illustrate our approach. In Chapter 7 we present a far more detailed analysis of a set of radiocarbon determinations where the robustness techniques advocated here are used within a complex model.

6.1.1 Outliers in radiocarbon dating

In Chapter 2 we discussed the reliability of radiocarbon dating. There we pointed out that outliers, that is, mistaken or discordant determinations, appear with relatively high frequency in radiocarbon dating. Baillie (1990) made an interesting study of the reliability of radiocarbon determinations. He analysed a series of radiocarbon analyses performed in an interlaboratory study involving 30 laboratories. Using dendrochronology, Baillie (1990) obtained the correct calendar age for a series of samples radiocarbon dated by the laboratories. He concludes that, for determinations taken until the early 1980's, about one-third were highly biased, (Baillie 1990)

"(...) the most damning for routine radiocarbon analysis is the 34% [of determinations] outside 200 radiocarbon years. From this, it would seem that a full one-third of all radiocarbon dates are effectively useless from the point of view of tight chronological research."

As Baillie explains, laboratories are continuously trying to improve their procedures and have, indeed, succeeded in providing better results. However, laboratories are not solely responsible for the quality of radiocarbon analyses. We can mention four main sources of error, three of which are outside control of the laboratories, which can affect the determinations and lead to the production of outlying radiocarbon determinations.

- (a) The quality of the care taken in the archaeological sampling in order to ensure that the samples can realistically provide results for the events we wish to date (see Bowman 1990 chapter 5).

- (b) The quality of the care taken in handling the samples on site in order to avoid contamination with older or younger material (see Bowman 1990 p. 27-28).
- (c) The quality of the care taken in sample handling and preparation in the laboratory to ensure that the samples undergo appropriate pretreatment and do not become contaminated before analysis for radiocarbon can take place (see Bowman 1990 p. 28-30).
- (d) Other non-controllable random factors that can appear during the whole process (see Scott *et al.* 1990 and 1990b).

The radiocarbon laboratories have only some control over factor (c) and no control over (a), (b) or (d). This means that radiocarbon laboratories can be producing top quality determinations with good estimates of the dating errors and still these determinations may be erroneous (outliers) since they may be discordant with

- (i) prior knowledge about the dating problem under study
- and
- (ii) other determinations available for the dating problem.

The outlier problem is not exclusively confined to radiocarbon laboratories and the correct assessment of the radiocarbon age and standard error involved in the dating process. On the contrary, it is a problem that involves both the laboratories and, specially, the users of radiocarbon since they are to suffer the most from the consequences of mistaken determinations.

Broadly speaking we estimate that, given the improved quality control protocols followed by laboratories since the late 1980's, at the present time 1 in 10 of all radiocarbon determinations obtained are discordant or mistaken and thus may be considered outliers. From this we see that of prime concern for archaeologists is the robustness of any statistical method used with a set of radiocarbon determinations to the presence of outliers.

At this point we envisage two questions that need attention.

- (i) How robust is our statistical framework to the presence of outliers?
- (ii) How can we improve our framework to allow for the possibility of outliers?

In the next Section we will analyse (i) through an example. It will not be difficult to notice the potentially unsatisfactory response of our models to the presence of outliers and the necessity of improving the robustness of our framework. Then in Section 6.2 we will present a general discussion of outliers in statistics in order to have a better understanding of the problem within the context of radiocarbon dating. From that, in Section 6.3 we will then consider (ii) where we formally define what we understand by an outlier and introduce a novel approach to the problem within the context of radiocarbon dating.

6.1.2 The summarisation problem: robustness to outliers

In Section 4.2.5 we studied the summarisation problem in which we were interested in the time-span and absolute position on the calendar scale of an archaeological *phenomenon*. Having a set of radiocarbon determinations $y_j \pm \sigma_j$'s associated with the calendar years θ_j 's we proposed $\psi = (\alpha, \beta)$ and

$$\theta_j \mid \alpha, \beta \sim U(\alpha, \beta),$$

and thus β and α represent the boundaries on the calendar scale for the time-span of the archaeological *phenomenon* under study. Then in Section 5.1 we studied a specific summarisation problem using data from the 'Chancay culture' in Peru (reported by Pazdur and Krzanowski 1991), using the prior knowledge that $\alpha > 400$ BP.

In other examples presented in Chapter 4 we used the above idea of having $\theta_j \mid \alpha, \beta \sim U(\alpha, \beta)$, commonly combined with other groups of determinations and more detailed prior information about the θ_j 's and the boundaries α and β . For example, we studied complicated archaeological phase structures related to groups of determinations, having

$$\theta_{i,j} \mid \psi_{2i-1}, \psi_{2i} \sim U(\psi_{2i-1}, \psi_{2i})$$

and different types of prior information about the ψ_i 's. The modelling used in the summarisation problem then represents a basic idea that forms part of several other more complicated modelling approaches. Therefore, it is of prime interest to analyse the robustness of this basic model, expecting this will give us an idea of the robustness of our framework in general.

To demonstrate the robustness to the presence of outliers of the modelling used in the summarisation problem, we revisit the Chancay culture dating problem, but now we introduce a fake, outlying, determination to the set of radiocarbon determinations available (see Table 5.1 in Section 5.2). The fake determination is 1400 ± 70 which is some 400 years away from the main bulk of the other determinations. The resulting marginal posterior distributions of α and β can be seen in Figure 6.1. The marginal posterior distributions obtained before, without the fake determination, are reproduced in the same Figure for comparison. We note that there is a substantial difference in the distribution of β from the original

distribution obtained before. With the original sample the 95% HPD region for β is 1200 to 940 BP, whereas including the fake determination the same region changes to 1400 to 1080 BP.

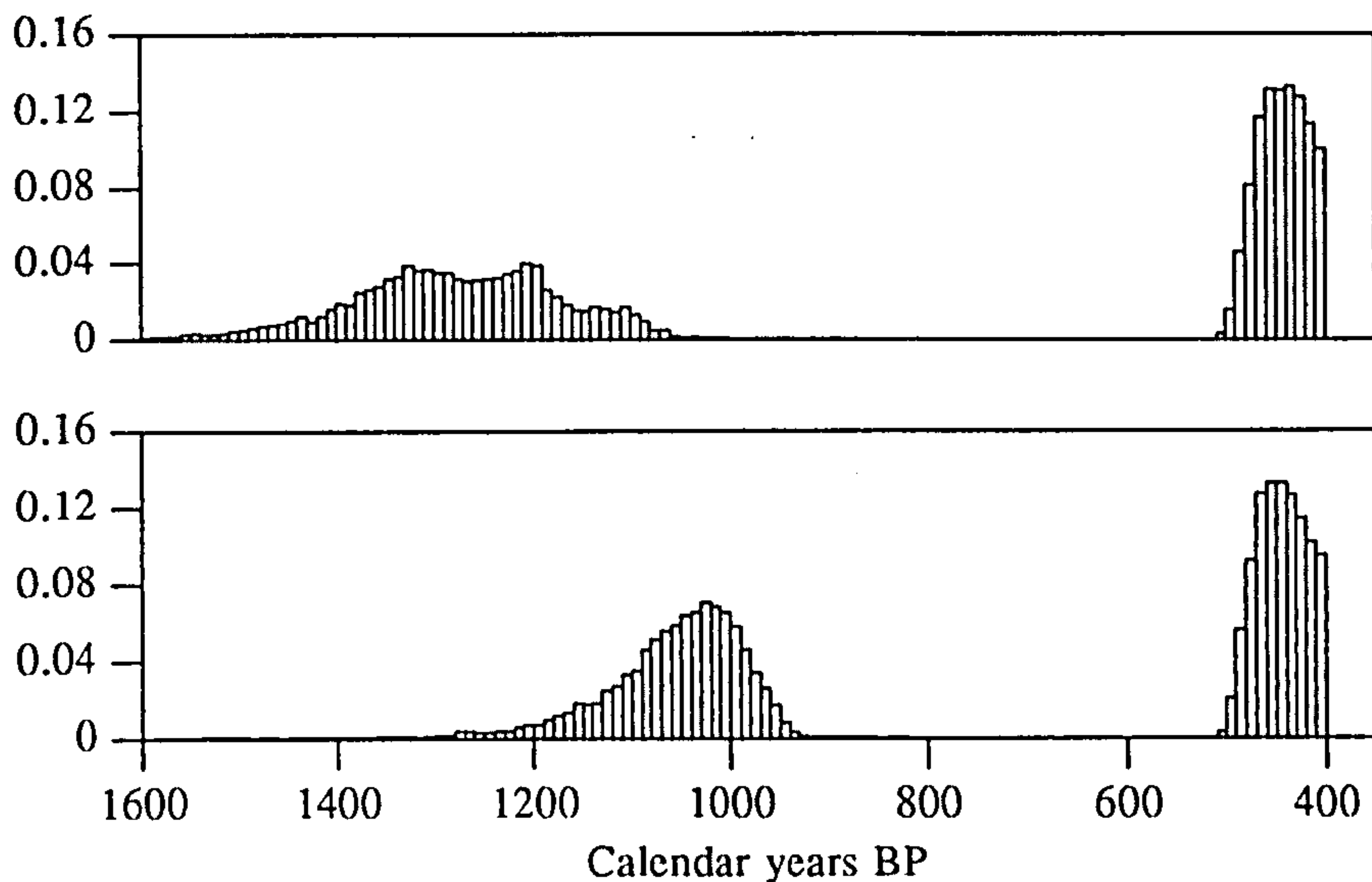


Figure 6.1

Histograms for the posterior distributions of β (left) and α (right), for the Chancay culture summarisation problem, with the original sample (below) and with the fake outlier determination 1400 ± 70 (above)

This example tells us that the presence of a single erroneous, outlying, radiocarbon determination can affect the whole of our posterior distributions and therefore all conclusions drawn from them. Indeed, the extent to which an outlying determination affects our inferences will depend on the model and prior information used for each problem. In this example, and since we are estimating the beginning and end of the Chancay culture, the (fake) evidence of an outlying, early, determination will affect our posterior knowledge about the beginning, β , of the culture. Note, however, that an outlying, late, determination cannot affect substantially our posteriori knowledge about α since we have the prior knowledge that $\alpha > 400$ BP.

The above example gives us evidence that our techniques might have an unsatisfactory response to outliers, with the potential danger of giving us non-robust inferences. Since, as indicated above, outliers are common in radiocarbon dating, we should consider them in our analyses in order to be able to make robust and reliable inferences.

Before we do so, we review and criticise a former approach to the problem of outliers in radiocarbon dating (the problem of outliers in radiocarbon dating has not been studied before using the Bayesian framework).

6.1.3 Previous approaches

At present there is no widely accepted technique or methodology to tackle the problems of outliers in radiocarbon dating. However, it is common practice in the interpretation of radiocarbon determinations to remove determinations as outliers, normally relying on informal and heuristic considerations. In the case of replications (a series of determinations considered to be associated with the same calendar year), Ward and Wilson (1978) proposed a statistical test to test 'that the series of determinations are consistent' (we reviewed this work in Section 2.4.3). Although this test was not designed to identify outliers, it is sometimes used for that purpose in mind. The test statistic is

$$T = \sum_{j=1}^m \frac{(y_p - y_j)^2}{\sigma_j^2},$$

where y_p is the pooled mean (see Equation 2.1). Under the assumption that all determinations have the same radiocarbon age, T has a chi-square distribution with $m-1$ degrees of freedom. If the resulting T is large and 'significant' (at some significance level), then the determinations are considered not to be consistent and the presence of an outlying radiocarbon determination in the sample is

suspected.

Although the T -test cannot be used to identify outliers, in a later paper Wilson and Ward (1981) develop a procedure that, based on statistic T , groups the determinations from a given sample to identify outliers. However, we have not found any reported applications of the techniques described in this paper. Below we discuss Wilson and Ward (1981) in greater detail.

They divide their study in three different cases namely, Case I, Case IIa and Case IIb. In Case I we have a sample of replicated determinations (replications). In Case IIa we do not have replicated determinations but expect a similar radiocarbon age for the determinations. Case IIb is the same as Case IIa but a 'sampling error' is introduced.

Given a set of radiocarbon determinations $y_1 \pm \sigma_1, y_2 \pm \sigma_2, \dots, y_m \pm \sigma_m$, for Case I samples Wilson and Ward (1981) consider the model

$$Y_j = \mu_j + e_j$$

where $e_j \sim N(0, \sigma^2)$. They propose to test the null hypothesis

$$H_0 : \mu_j = \mu \quad (j = 1, 2, \dots, m)$$

against the alternative H_1 that μ_j is equal to ϕ_1 or ϕ_2 (with at least one μ_j equal to either values), where ϕ_1 and ϕ_2 are the radiocarbon ages of two different groups. In other words, we are testing the hypothesis that the sample represents one group with a single radiocarbon age against the alternative that the sample represents two groups with different radiocarbon ages.

To test H_0 Wilson and Ward (1981) order the determinations according to their radiocarbon age and then use the test statistic $\Delta = \max \Delta_k$ where

$$\Delta_k = T_{[1,m]} - T_{[1,k]} - T_{[k+1,n]}.$$

Here $T_{[p,q]}$ is the T statistic given above but only considering the elements p to q of the ordered sample. Wilson and Ward give 95% upper quantiles for the distribution of Δ under H_0 to perform the test. They explain that, if H_0 is rejected (at some significance level), the sample may be split into two groups containing the determinations y_1 to y_{k^*} and y_{k^*+1} to y_m , of the ordered sample, where $\Delta = \Delta_{k^*}$. The test can then be applied to subgroups of the sample and obtain subgroups of determinations 'judged to be homogeneous by applying the Δ -criterion', at some significance level. It is then *re-assured* that if an outlier is present in the sample, it will form a group on its own.

For Case IIa samples Wilson and Ward (1981) propose the model

$$Y_j = \mu_j + e_j + f_j + g_j$$

where $e_j \sim N(0, \sigma_j^2)$, $f_j \sim N(0, \sigma_f^2)$ and $g_j \sim N(0, \sigma_g^2)$ are normal errors (commonly taken to be independent) that account for the error arising from calibration (commonly $\sigma_f = 50$ and $\sigma_g = 70$, see Section 2.4.3). No explicit calibration method is used. They propose to test H_0 as above using basically the same test statistic Δ as in Case I samples and, if necessary, proceed to split the sample in subgroups.

Note that in the above two cases, it is assumed that the sample has a single radiocarbon age, μ , under the null hypothesis H_0 (or within each group, ϕ_1 or ϕ_2 , under H_1). That is, we expect that the only variation in the sample is due to the standard errors σ_j 's and, in Case IIa, to errors arising from calibration. Wilson and Ward (1981) explain that in most archaeological examples this cannot be assumed and a 'sampling error' must be considered. For example, we might be sampling from a specific stratigraphic layer that could have lasted several hundreds years and, therefore, we cannot assume that the samples have the same radiocarbon age. They then propose the following model for Case IIb samples

$$Y_j = \mu_j + r_j + e_j + f_j + g_j$$

where e_j, f_j, g_j are defined as above and $r_j \sim N(0, \sigma^2)$ is the sampling error (on the radiocarbon scale). Here σ^2 is unknown and must be estimated from the determinations. They then propose to test H_0 as above (in this case the test statistic is more complex although similar to Δ), and split the sample if necessary.

To illustrate their techniques we reconsider the Chancay culture data (studied in the previous Section) with the fake outlying determination 1400 ± 70 . The dating problem here is to estimate the beginning, β , and the end, α , of the Chancay culture. We know that the objects radiocarbon dated all belong to the Chancay and from that we wish to estimate α and β (in the previous Section we used the prior knowledge that $\alpha < 400$ BP). Since the objects dated all belong to the Chancay culture, that could have lasted several hundreds of years, we cannot assume a single radiocarbon age and, therefore, we have a Case IIb sample. That is, we need to include sampling error r_j in our model.

However, using Wilson and Ward (1981) techniques it is not clear how to introduce into the analysis the characteristics of the dating problem under study. That is, it is not clear how to estimate the beginning and end of the Chancay culture and identify outliers. The Wilson and Ward Δ -criterion is fixed and cannot adapt to suit the various situations that could appear. They seem to recognise the limitations of their techniques and explain that

"It must be stressed that (...) [the] assumption of normality for r_j may not be particularly realistic but we have made it for comparative mathematical simplicity. It is not easy to decide the form of the actual true age distribution in real time of the site or strata or series of sites from which we are sampling, let alone the conversion of this actual age distribution in real time to the distribution in radiocarbon years b.p!"

We believe that the sampling error term r_j cannot be used to properly model the characteristics of the dating problem under study. Thus Wilson and Ward (1981) techniques can hardly be used for this example. However, to illustrate their technique, we applied their Δ -criterion to the Chancay data set including the fake outlying determination 1400 ± 70 . Firstly, we performed the test without sampling error (Case IIa) and, at 95% significance level, split the sample into 5 subgroups. Here the fake determination is contained in a group of its own. Secondly, we consider sampling error and, at 95% significance level, there was no evidence to split the sample (including the fake determination). As a further illustration, in Figure 6.2, we present these results, together with our estimates for α and β . In any case, we must stress the fact that grouping the determinations does not solve the dating problem at hand.

As a general critique to Wilson and Ward (1981), we make the following remarks. As explained before in Section 2.4.3, the Ward and Wilson's (1978) T -test has a series of inadequate characteristics arising from the fact that no calibration method is used. Grouping radiocarbon determinations according to their T values, which is the basic idea behind Wilson and Ward (1981), cannot be an acceptable procedure for the identification of outliers since,

(i) no calibration procedure is used

and

(ii) no archaeological considerations can be used in the analysis.

Point (i) is itself a sufficiently important issue that alone questions the validity of their procedure. Moreover, we believe that (ii) represents a crucial factor in the correct identification of outliers. As seen in previous Chapters, the models used for the analysis and interpretation of radiocarbon determinations vary

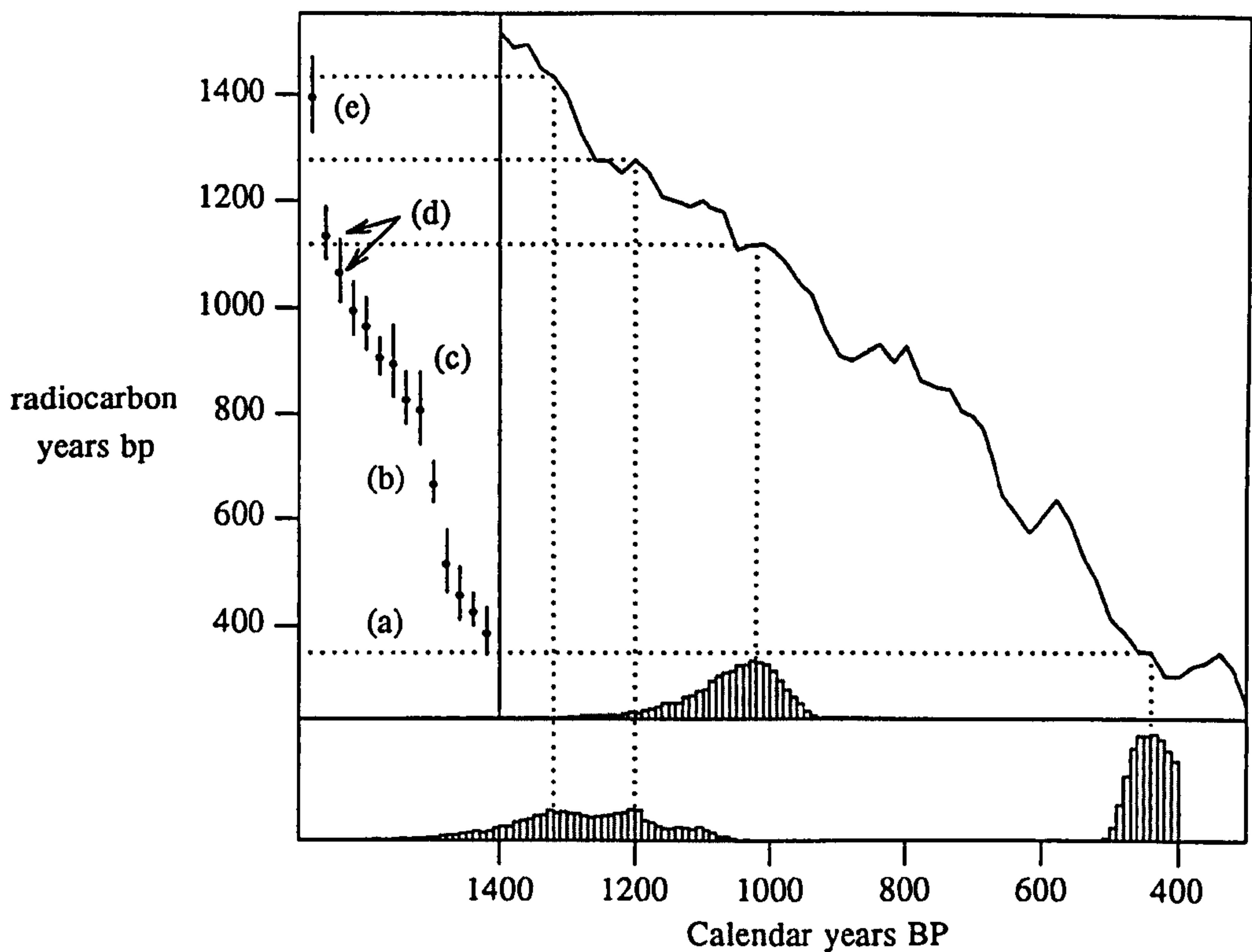


Figure 6.2

Radiocarbon determinations for the Chancay culture (left) and histograms for the posterior distributions of α (right) and β (left), considering and not considering the outlying determination 1440 ± 70 . We also present the posterior modes for α and β and their corresponding radiocarbon ages. The determinations are divided into groups (a) to (e) according to Wilson and Ward (1981), Case IIa.

depending on the archaeological dating *phenomena* under study. For this reason the judgement of what is or what is not an outlier *must* depend on the specific model used to explain the determinations and thus, ultimately, depends on the archaeological *phenomena* of interest. We must then seek statistical techniques that allow us to include archaeological considerations into the analysis and, in addition, are robust to the presence of outliers.

To our knowledge, no other technique for the identification of outliers in radiocarbon data has been proposed. In the remainder of the Chapter we will try to develop a methodology to identify outliers in radiocarbon data.

6.2 Outliers in statistics

The detection of outliers and the development of methods that are robust against outliers are very important topics in statistics. In general outliers are data points that have unexpected or exceptional values. As explained before, in this Section we will try to present the problem of outliers in radiocarbon dating within the global perspective of outliers in statistics. This will, hopefully, give us insight into what approach should be followed to improve the robustness of our framework to the presence of outliers.

Barnett and Lewis (1978, 1984) and Barnett (1978) have studied the problem of outliers from a global perspective. Barnett (1978) points out that, before any specific technique or method can be considered, the outlier problem at hand should be put in the perspective of three

"(...) fundamental interrelated questions:

- (i) what are the possible causes of outliers in statistical data?
- (ii) in what way do outliers influence data analysis, OR what are our different possible aims at processing outliers?
- (iii) what probabilistic models might be employed to explain the presence of outliers?"

We will briefly discuss (i) to (iii) in turn.

With respect to the possible causes for the generation of outliers, Barnett (1978) points out that they can be classified within three groups, namely:

- (i) The 'inherent' causes. These produce extreme or extraordinary data points but that, given the characteristics of the specific problem under study, are expected to occur with some frequency and are considered an inherent part

of the problem modelled.

- (ii) The 'measurement causes'. These represent mistakes in the production of the data, from a mislocation of a decimal point to malfunctions in measurement instruments.
- (iii) The 'execution causes' which are mistakes taken in the sampling process, specifically, the inclusion in the sample of an individual not belonging to the sample space.

Once we have identified the causes for outliers in a particular statistical problem, we should decide our aims at processing outliers. Indeed, this will depend on the causes identified for the generation of outliers. In broad terms we can distinguish two general aims, namely,

- (i) to 'accommodate' outliers with the rest of the data

and

- (ii) to 'identify' outliers, and question their presence and validity, within the rest of the data.

In general, if we have identified 'inherent' causes for the generation of outliers, this suggests that we should aim to *accommodate* them within the rest of the data. This is commonly carried out using robust techniques that explain the data as a single whole. As opposed to this, if we suspect 'measurement' or 'execution' causes, this suggests that we should aim to *identify* outliers and question their presence within the rest of the data. Commonly this is carried out using a principal or central model that explains most of the data and alternative models that explain outliers.

Finally we should consider a suitable probabilistic model to explain the presence of outliers in a particular problem. This, of course, will be dependent on the causes identified for the presence of outliers and the aims at processing them. In a very general way one can mention three modelling approaches that have been used to explain outliers in statistical data, namely:

- (i) The 'inherent' modelling approach. In this case all data $y_j \sim F$, where F explains both outliers and non-outliers. This approach is commonly appropriate to accommodate outliers within the rest of the data.
- (ii) The 'mixture model' approach. In this case, given that y_j is not an outlier

$$y_j \sim F$$

where F represents the central model that explains most of the data, whereas if y_j is an outlier

$$y_j \sim G$$

where G represents the alternative model that explains all outliers. This approach models the expectation that some (few) data points were generated by a different process (modelled with G) than the rest and is used to identify outliers (those data points better explained with model G).

- (iii) The 'slippage model' approach. This model states that, given that y_j is not an outlier,

$$y_j \sim F$$

where F represents the central model, and given that y_j is an outlier

$$y_j \sim F_j$$

where F_j represents an alternative model that explains the presence of y_j as an outlier. In this case we are trying to model the presence of mistaken data points, each one generated by a particular (extraordinary) process modelled with F_j . This approach is used to identify outliers (those better explained with an alternative model F_j).

Several authors have considered the above modelling approach to tackle the problem of outliers in statistical data. Within Bayesian statistics we can mention Dawid (1973), Hill (1974), O'Hagan (1979), among others, that have studied 'inherent' modelling approaches to accommodate outliers within the rest of the sample. In contrast to this, mixture and slippage models have been proposed for the identification of outliers by, among others, Box and Tiao (1968), Abraham and Box (1978) and Guttman *et al.* (1978). These latter models can be described as follows.

Box and Tiao (1968) propose a 'scale-inflated mixture model' for outliers in normal samples. That is, given that y_j is not an outlier

$$y_j \mid \sigma^2 \sim N(0, \sigma^2)$$

and given that y_j is an outlier

$$y_j \mid b, \sigma^2 \sim N(0, b \sigma^2),$$

for $b > 1$. Here outliers are thought to be generated by a (single) alternative model that has a higher variance than the model generating the rest of the sample.

Abraham and Box (1978) propose a 'location-shift mixture model' for outliers in normal samples. The model proposed is

$$y_j \mid \mu, \delta, \phi_j, \sigma^2 \sim N(\mu + \delta \phi_j, \sigma^2)$$

where the random variable ϕ_j is defined as

$$\phi_j = \begin{cases} 1 & \text{if } y_j \text{ needs a shift} \\ 0 & \text{otherwise.} \end{cases}$$

We then interpret ‘ y_j needs a shift’ as ‘ y_j is an outlier’. Here then, if $\phi_j = 0$, y_j is explained with the model $N(\mu, \sigma^2)$ and if $\phi_j = 1$, y_j is explained with the ‘shifted model’ $N(\mu + \delta, \sigma^2)$.

Guttman *et al.* (1978) propose a ‘location-shift slippage model’ for outliers in normal samples. Their model states that, given that y_j is not an outlier,

$$y_j \mid \mu, \sigma^2 \sim N(\mu, \sigma^2)$$

and, given that y_j is an outlier,

$$y_j \mid \mu, \delta_j, \sigma^2 \sim N(\mu + \delta_j, \sigma^2).$$

Here outliers are explained by (several) alternative ‘shifted’ models $N(\mu + \delta_j, \sigma^2)$. As opposed to Abraham and Box’s (1978) approach where only one type of ‘shifted’ model is allowed, in this case various shifts are considered. Thus a sample may have an upper and a lower outlier at the same time (positive and negative δ_j ’s, respectively).

Using this general discussion on outliers in statistics we will consider the problem in radiocarbon data. This we do in the next Section.

6.2.1 What can we do?

In this Section we discuss a suitable approach to the problem of outliers in the analysis and interpretation of radiocarbon data. In turn we will discuss (i) the causes of outliers, (ii) our aims at processing outliers and (iii) the models we can use to explain the presence of outliers in radiocarbon data. Our approach will then be based on this discussion and developed in later Sections.

It is indeed worrying for us to note that at least two of the three general causes for the production of outliers mentioned in the previous Section, can be observed in the radiocarbon dating process, namely:

- (i) Measurement causes. Even with the best quality control protocols, the production of outlying radiocarbon determinations due to the chemical and physical processes involved cannot be discounted. The complex and sophisticated processes involved in radiocarbon dating are indeed error prone and not even the best laboratories in the world could guarantee perfection. Furthermore, contamination on site due to improper handling of samples by archaeologists can lead to misleading determinations. Thus factors (b), (c) and (d), mentioned in Section 6.1.1, can be classified as measurement causes for the generation of outliers in radiocarbon data.
- (ii) Execution causes. The sampling process is always a factor of error in archaeology. Occasionally samples taken from excavations are mistakenly related to contexts they do not belong to. This is a result of both misjudgements by archaeologists in the excavation process and by truly unforeseeable circumstances. We then classify factor (a), mentioned in Section 6.1.1, as an execution cause for the generation of outliers in radiocarbon data.

Therefore we can say that we expect outliers in radiocarbon dating to be related both to measurement and execution errors. Moreover, in comparison to other statistical contexts, outliers in radiocarbon dating are far more common (we mentioned before that we would expect 1 in 10 determinations to be highly biased).

Why is it then that we continue to call 'outliers', data points known to be affected by measurement and execution errors with relatively large frequency?

Should we continue to understand outliers in radiocarbon dating as 'erroneous' or 'mistaken' determinations or perhaps view them as 'correct and extreme' data points?

The answers to the above questions are not complicated. Provided we can trust our model and prior information about the dating problem under study, outliers will be erroneous determinations independently of their frequency. The statistical analysis of sets of radiocarbon determinations represents a very special case in which there is substantial or even abundant prior information about the processes involved in the problem. Even if we have not established a formal probabilistic model, the samples submitted for a radiocarbon analysis are, in most of the cases, related to archaeological contexts that provide (at least broad) information about the age of the samples. A sample submitted for radiocarbon dating cannot have a completely arbitrary age.

For example, we could obtain a radiocarbon determination with 'modern ^{14}C activity' (less than 50 years of age) or a determinations with 'background ^{14}C activity' (more that 30,000 years of age). For the vast majority of archaeological contexts such determinations are discordant with the dating problem under study and it is common practice to reject them as erroneous, excluding them from the subsequent analysis and interpretation. From the above we conclude the following.

Outliers in radiocarbon dating are discordant and mistaken determinations that appear with (unfortunately) relatively large frequency and are related to measurement and execution errors. With this respect, we wish then to identify discordant determinations (outliers), present in radiocarbon data and, if necessary, remove them from subsequent analysis and interpretation.

However, it is not clear when can we identify a radiocarbon determination as 'discordant' apart from the obviously extreme cases of modern and background ^{14}C activity. An explicit statistical approach will be needed to measure the accordance (or discordance) of radiocarbon determinations with the model, prior information and other determinations available in order to identify outliers in a specific dating problem.

In the previous Section we mentioned two general modelling approaches to identify outliers, namely, the mixture and slippage models. For the case of radiocarbon dating, we do not have information about a particular type of process that would generate *all* outliers, as is the case for a mixture model. We would expect any sort of extreme determinations (high or low radiocarbon age) and, in principle, a single set of determinations can have both upper and lower outliers at the same time. With this in mind we believe that the slippage modelling approach is the most appropriate to model the presence of discordant radiocarbon determinations. That is, discordant determinations will be those better explained with a 'discordant model' F_j than with the central model F . In this way any sort of discordancies, appropriate to each determination, can be modelled. This is the approach we will try to implement here.

In addition, we must remember that one of the basic principles of the Bayesian framework is that uncertainty should be measured with a probability distribution. Undoubtedly there will be a degree of uncertainty about the accordance or discordance of a particular determination with the rest of the factors in a given dating problem. That is, there will always be a degree of uncertainty as to when a determination can be considered an outlier or not. Therefore, in order to measure that uncertainty with a probability distribution our approach ought to formally define the event that any 'determination y_j is an outlier'.

We mentioned in the previous Section mixture and slippage models, namely, the scale-inflated and the location-shift mixture models and the location-shift slippage model. Certainly, the idea of modelling discordant radiocarbon determinations with different degrees of location-shifts seems very appealing. That is, if y_j is not discordant, $y_j \sim N(\mu, \sigma_j^2)$ and if y_j is discordant $y_j \sim N(\mu + \delta_j, \sigma_j^2)$. Therefore we should explore an approach similar to Guttman's *et al.* (1978) location-shift slippage model. However, we want to formally define the event that any 'determination y_j is an outlier'. In this respect, Abraham and Box's (1978) location shift mixture model approach is the more appealing since they explicitly define such an event when introducing the random variable ϕ_j . That is, if $\phi_j = 0$, $y_j \sim N(\mu, \sigma_j^2)$ and if $\phi_j = 1$, $y_j \sim N(\mu + \delta, \sigma_j^2)$, thus y_j is an outlier if $\phi_j = 1$. (We will introduce calibration and further details in the next Section.)

An obvious alternative would be to inflate the variance of each determination, a similar idea to Box and Tiao's (1968) scale-inflated model. That is, if y_j is an outlier $y_j \sim N(\mu, b \sigma_j^2)$ for $b > 1$. In principle, this approach seems as valid as the location-shift approach. However, we believe that the scale-inflated option does not properly describe the general outlier generation process present in radiocarbon data.

If we inflate the variance we would examine for changes in the *precision* of each determination. That is, we would doubt about the validity of the standard error σ_j reported by the laboratories and, in fact, we would attribute the generation of outliers to changes in the variances σ_j^2 . This is not precisely what we are trying to do since what we are trying to examine is the overall *accuracy* of the determinations and not only their precision.

For example, suppose that a sample submitted for radiocarbon dating is expected to have a radiocarbon age of around 600 radiocarbon years bp and the determination reported turns out to be 1600 ± 50 bp. Suppose that the reason for

this discordancy is that the sample *was thought* to belong to some context although, for unfortunate circumstances, it did not (of course, this represents an 'execution' error that was not suspected before submitting the sample for dating). The radiocarbon determination reported may still correctly represent the radiocarbon age for the sample dated and, therefore, the determination may still be precise.

An increase in the variance 50^2 cannot alone explain the discordancy and therefore a shift on the radiocarbon scale should be suspected.

We therefore intend to use a combination of the modelling approaches suggested by Abraham and Box (1978) and Guttman *et al.* (1978) to tackle the problem of outliers in radiocarbon dating, that is, a location-shift slippage modelling approach. This we do in the next Section.

6.3 Our approach

6.3.1 Shift outliers

The aim now is to improve our general methodology in order to have an explicit and satisfactory way to deal with outliers in radiocarbon dating. For the reasons explained above, we propose a combined version of the approaches first presented by Abraham and Box (1978) and Guttman *et al.* (1978). Suppose that we have a set of radiocarbon determinations $y_1 \pm \sigma_1, y_2 \pm \sigma_2, \dots, y_m \pm \sigma_m$ associated with the (unknown) calendar years $\theta_1, \theta_2, \dots, \theta_m$. The basic idea then is to state that y_j is an outlier if y_j needs a shift δ_j (on the radiocarbon scale) in order that it be in accordance with the rest of the sample.

We formalise this by stating that

$$y_j \mid \theta_j, \delta_j, \phi_j \sim N(\mu(\theta_j) + \phi_j \delta_j, \omega_j^2(\theta_j))$$

($\omega_j^2(\theta_j) = \sigma_j^2 + \sigma^2(\theta_j)$), where

$$\phi_j = \begin{cases} 1 & \text{if } y_j \text{ needs a shift} \\ 0 & \text{otherwise} \end{cases}$$

and we interpret ‘needs a shift’ as ‘is an outlier’. In other words, a mistake in the radiocarbon dating process will be modelled by a shift in y_j from the true radiocarbon year $\mu(\theta_j)$. To illustrate this, in Figure 6.3, we present the hierarchy diagram for the extended basic model $y_j \mid \theta_j, \delta_j, \phi_j$.

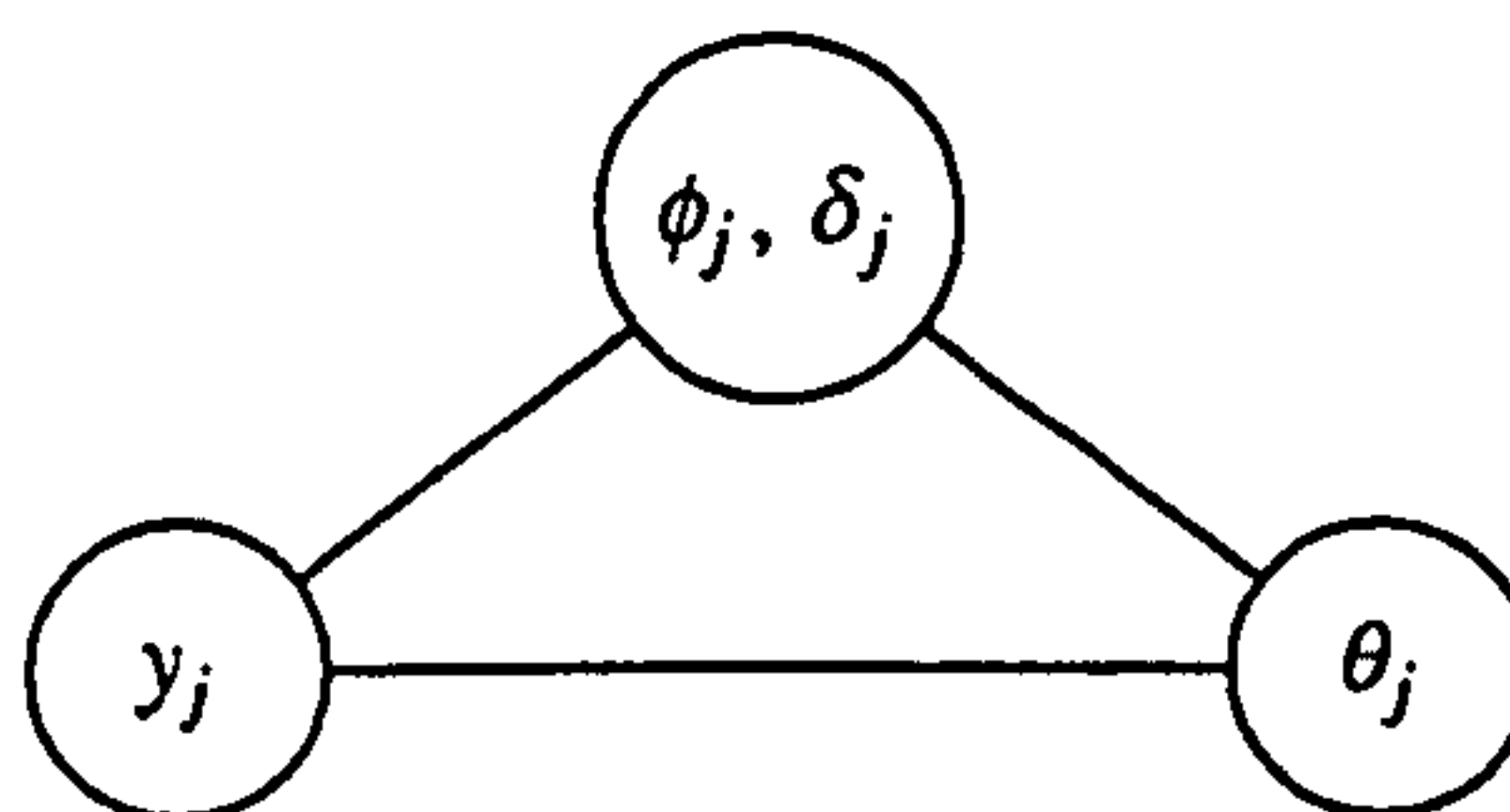


Figure 6.3

Hierarchy diagram for the extended basic model.

We assume ϕ_j and δ_j to be independent, a vague prior distribution for δ_j and a prior probability

$$P[\phi_j = 1] = p_j$$

which represents the prior belief that determination y_j is an outlier. In general, if a sample is believed, *a priori*, to be affected by one of the factors (a)-(d) (explained in Section 6.1.1), this could be reflected in a bigger p_j for the corresponding determination. If we are to be ‘fair’ in our analysis we *must* state p_j *before* the radiocarbon determination is seen and not, for example, give a higher p_j to a determination that looks, *a posteriori*, like an outlier from

examination of the data. To avoid such a misleading usage of this methodology we could state a single $p = p_j$ for all $j = 1, 2, \dots, m$ before any determination is observed and thus be sure that the prior distribution of each ϕ_j is indeed *a priori*.

The key idea of this methodology is to obtain the posterior distributions of each ϕ_j , that is, $P[\phi_j = 1 \mid y]$. In other words, the posterior probability that y_j is an outlier. To calculate $P[\phi_j = 1 \mid y]$ we will need to include in the analysis the parameter ψ from our framework and define $f(\theta \mid \psi)$ and $f(\psi)$ to introduce the characteristics of the dating problem under study (we discuss this issue in the next Section). Therefore, the posterior probability $P[\phi_j = 1 \mid y]$ that y_j is an outlier will not only be based on the data y . It will consider

(i) the calibration curve and the variance in the calibration curve (since we are using $\mu(\theta)$ and $\sigma^2(\theta)$ in our model)

(ii) the relationship between the objects radiocarbon dated and the dating *phenomena* of interest (introduced in the analysis via $f(\theta \mid \psi)$)

and

(iii) further prior information about the *phenomena* under study (described by $f(\psi)$).

To summarise, what we identify as an outlier will very much depend on the dating problem considered. Below we formalise these ideas.

6.3.2 The extended framework

According to the framework presented in Chapter 4 we need to define ψ , the model $f(y \mid \theta)$ and the prior distributions $f(\theta \mid \psi)$ and $f(\psi)$. The initial step is to

define ψ according to the specific dating *phenomena* under study. In Chapter 5 we have already presented various examples of this. As discussed above, our approach to identify outliers is to use the model

$$y_j \mid \theta_j, \delta_j, \phi_j \sim N(\mu(\theta_j) + \phi_j \delta_j, \omega_j^2(\theta_j))$$

(where $\omega_j^2(\theta_j) = \sigma_j^2 + \sigma^2(\theta_j)$).

Therefore, within this extended framework, the model (likelihood) is $L(y \mid \theta, \psi, \delta, \phi) = L(y \mid \theta, \delta, \phi)$ and, assuming that given the θ_j 's the determinations are independent, we have

$$L(y \mid \theta, \delta, \phi) = \prod_{j=1}^m L(y_j \mid \theta_j, \delta_j, \phi_j) \propto \prod_{j=1}^m \frac{1}{\omega_j(\theta_j)} \exp \left\{ -\frac{1}{2} \left(\frac{(y_j - (\mu(\theta_j) + \phi_j \delta_j))^2}{\omega_j^2(\theta_j)} \right) \right\}.$$

Now, when we define the distributions $f(\theta \mid \psi)$ and $f(\psi)$, the radiocarbon determinations, y , are not (and must not be) considered since these distributions are stated *a priori*. Thus, the specification of such distributions cannot be influenced by the shifts δ_j 's, on the radiocarbon scale, that might affect the determinations when these are observed. Outliers in radiocarbon dating arise as a result of several different factors, some of which might be produced on site by the archaeologists. However, the fact is that outliers manifest only on the radiocarbon scale through the values observed for the y_j 's. Thus outliers are modelled as shifts on the radiocarbon scale and, *a priori*, the relationship between θ and ψ is not influenced by such shifts. This means that $\theta \mid \psi$ is (and must be) independent of the δ_j 's and of the ϕ_j 's.

Therefore, given a specific dating problem the prior distributions $f(\theta \mid \psi)$ and $f(\psi)$ should remain the same whether we use the above methodology to allow for outliers or our former methodology discussed in Chapters 4 and 5. (This has

the immediate consequence that the methodology to identify outliers presented here can be applied to *all* of the examples presented in Chapter 5.)

Let $\delta = (\delta_1, \delta_2, \dots, \delta_m)$ and $\phi = (\phi_1, \phi_2, \dots, \phi_m)$. From the above discussion we can see that the prior distributions $f(\theta | \psi)$ and $f(\psi)$ do not depend on δ nor on ϕ and therefore we can see that, for the full conditionals of the ψ_i 's,

$$f(\psi_i | y, \theta, \psi_{-i}, \delta, \phi) = f(\psi_i | \theta, \psi_{-i}) \propto f(\theta | \psi) f(\psi_i | \psi_{-i}).$$

These full conditionals are identical to the full conditionals for the ψ_i 's given before in Chapter 4. To illustrate this, in Figure 6.4 we present the hierarchy diagram for the parameters involved in this extended framework. As a result of this, the full conditionals for the ψ_i 's remain the same and, therefore, all the techniques and routines for sampling from them.

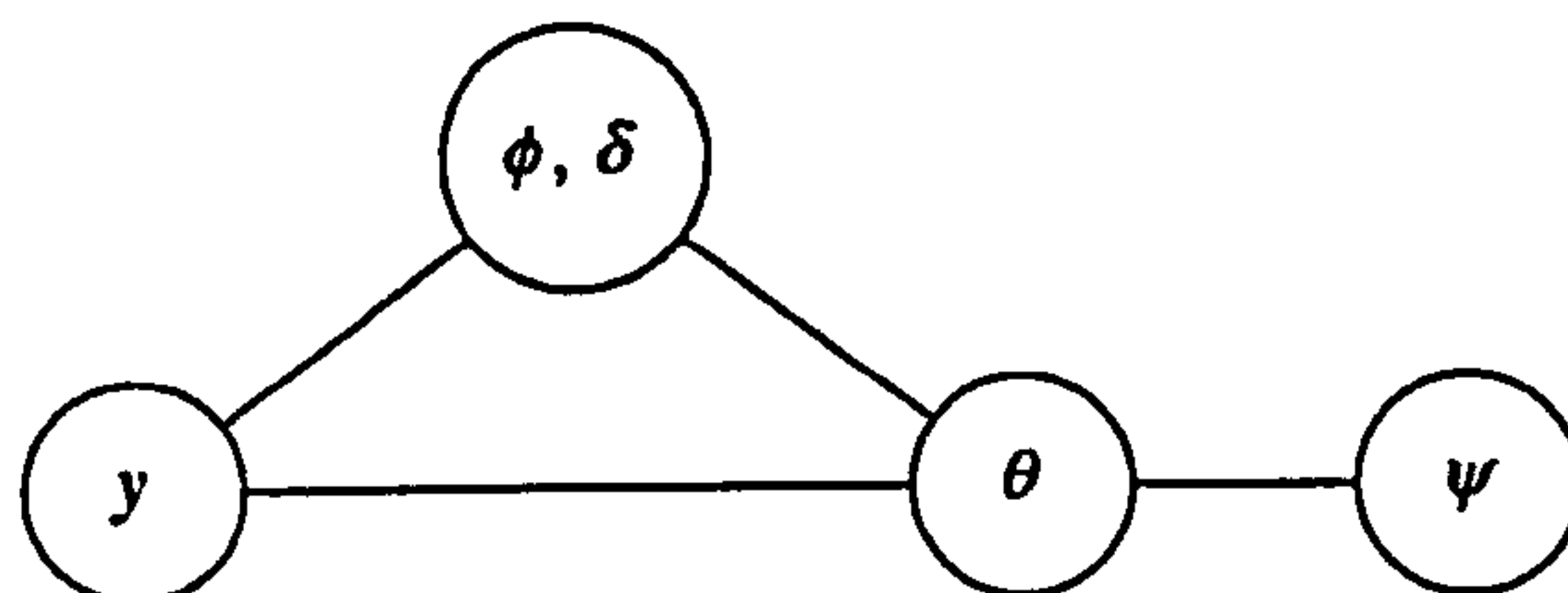


Figure 6.4
Hierarchy diagram for the extended framework.

The full conditionals for the θ_j 's do change and these are

$$f(\theta_j | y, \theta_{-j}, \psi, \delta, \phi) \propto \frac{1}{\omega_j(\theta_j)} \exp \left\{ -\frac{1}{2} \left(\frac{(\mu(\theta_j) - (y_j - \phi_j \delta_j))^2}{\omega_j^2(\theta_j)} \right) \right\} f(\theta_j | \theta_{-j}, \psi).$$

However, these distributions are the same as the ones used in Chapter 5 for sampling the θ_j 's, but now using at every step the shifted radiocarbon age $y_j - \phi_j \delta_j$ instead of y_j . Thus, in principle, the routines implemented before could be easily

modified to be used here.

Using the prior distributions

$$P[\phi_j = 1] = p_j \quad (6.1)$$

for the ϕ_j 's and assuming a vague prior for the δ_j 's, independent of ϕ , the rest of the full conditionals are

$$\delta_j \mid y, \theta, \delta_{-j}, \phi \sim N(y_j - \mu(\theta_j), \omega_j^2(\theta_j)), \text{ for } \phi_j = 1,$$

and

$$f(\phi_j \mid y, \theta, \delta, \phi_{-j}) \propto p_j^{\phi_j} (1-p_j)^{1-\phi_j} \exp \left\{ -\frac{1}{2} \left(\frac{(\mu(\theta_j) - (y_j - \phi_j \delta_j))^2}{\omega_j^2(\theta_j)} \right) \right\}.$$

Since the full conditionals for the δ_j 's are normal and for the ϕ_j 's are Bernoulli distributions (that need normalising), it is not difficult to sample from them. Using a MCMC method we will be able to obtain estimates of the distributions of $\phi_j \mid y$, thereby giving the probability of each determination being an outlier, given the current sample.

Below we present two examples to illustrate how this extended framework can be used. We reconsider the replications and the summarisation problems studied before.

6.4 Examples

6.4.1 Replications

We now reconsider the replications problem studied before in Section 4.4, in the light of our extended framework presented above. Let us suppose that we have a set of radiocarbon determinations $y_1 \pm \sigma_1, y_2 \pm \sigma_2, \dots, y_m \pm \sigma_m$ which are

associated with the same calendar year. That is, we are working with a set of replicated determinations and we have that $\theta_j = \theta$, for $j = 1, 2, \dots, m$, where θ_j is the associated calendar year for each determination. In that case our model is

$$y_j \mid \theta, \delta_j, \phi_j \sim N(\mu(\theta) + \phi_j \delta_j, \omega_j^2(\theta))$$

(where $\omega_j^2(\theta) = \sigma_j^2 + \sigma^2(\theta)$). As in Section 4.4, in this case the vector ψ from our general framework is not random and thus we concentrate our attention on the posterior distributions of θ , δ and ϕ . The full conditionals in this case are (assuming that the determinations are independent given θ)

$$f(\theta \mid y, \delta, \phi) \propto \prod_{j=1}^m \frac{1}{\omega_j(\theta)} \exp \left\{ -\frac{1}{2} \left(\frac{(\mu(\theta) - (y_j - \phi_j \delta_j))^2}{\omega_j^2(\theta)} \right) \right\} f(\theta),$$

$$\delta_j \mid y, \theta, \delta_{-j}, \phi \sim N(y_j - \mu(\theta), \omega_j^2(\theta)), \text{ for } \phi_j = 1,$$

and

$$f(\phi_j \mid y, \theta, \delta, \phi_{-j}) \propto p_j^{\phi_j} (1-p_j)^{1-\phi_j} \exp \left\{ -\frac{1}{2} \left(\frac{(\mu(\theta) - (y_j - \phi_j \delta_j))^2}{\omega_j^2(\theta)} \right) \right\},$$

where $f(\theta)$ is the prior distribution for θ .

To illustrate this application of our methodology we reconsider the set of radiocarbon determinations arising from the Shroud of Turin (see Table 4.1 in Section 4.4), now using our extended framework to investigate for outliers in the sample. Since all determinations arise from a single object (the Shroud) we can assume that we have a set of replications and thus $\theta_j = \theta$ as above. Furthermore, since the Shroud was first displayed in the 1350's (AD), we know that $\theta > 600$ BP.

For comparison, and before we proceed with our analysis, we applied the T -test and Wilson and Ward's (1981) Δ -criterion to the Shroud data (explained in Section 6.1.2). The T statistic is equal to 20.9 and therefore, with a 95%

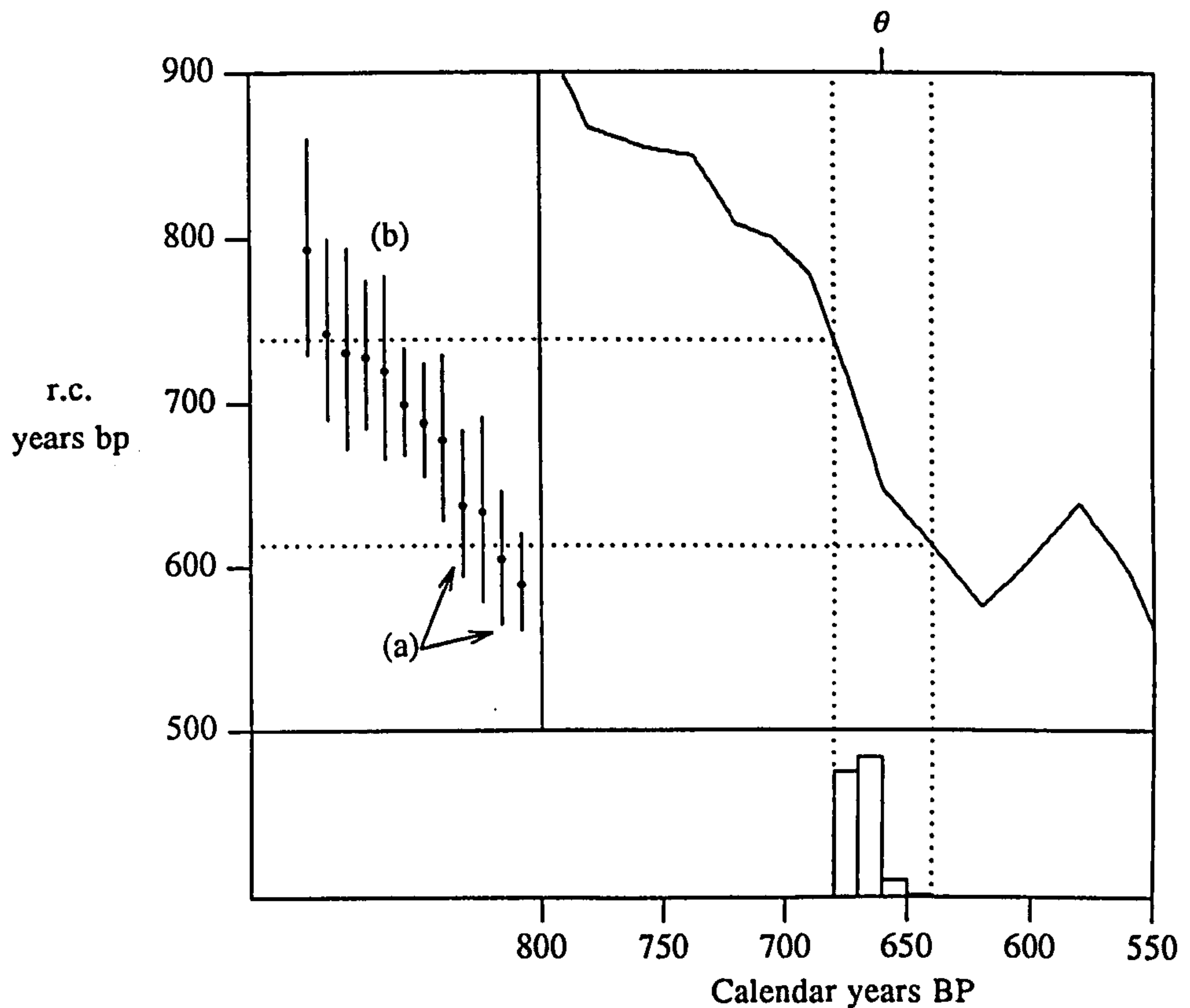


Figure 6.5

Radiocarbon determinations for the Shroud of Turin and our estimate for their associated calendar year shown as a histogram. The determinations were divided into subgroups (a) and (b), according to Wilson and Ward's (1981) Δ -criterion.

confidence level, the sample is 'not consistent'. We then applied Wilson and Ward's (1981) Δ -criterion, Case I, to cluster the determinations and obtained two subgroups using a significance level of 95%. The subgroups are shown in Figure 6.5. Note that the sample spans more than 200 radiocarbon years and, perhaps for this reason, it was rejected as having a single radiocarbon age, according to the T -test. As seen in Figure 6.5 two subgroups are then formed using the Δ -criterion. However, this picture will change substantially when we use our methodology, that includes calibration, and evaluates the problem on the calendar scale.

We take all 12 radiocarbon determinations and, using a single prior probability $p_j = 0.1$ for $j = 1, 2, \dots, 12$, calculated the posterior probabilities $P[\phi_j = 1 \mid y]$ using the Gibbs sampler. The results can be seen in Table 6.1, column (iv).

| Sample (rep.) | j | (j) | P_j | P_j |
|------------------|-----|-------|-------|-------|
| A1.1 | 1 | 1 | 0.23 | 0.20 |
| A1.2 | 2 | 6 | 0.11 | 0.11 |
| A1.3 | 3 | 2 | 0.12 | 0.11 |
| A1.4 | 4 | 7 | 0.15 | 0.17 |
| O1.1 | 5 | 12 | 0.35 | - |
| O1.2 | 6 | 9 | 0.21 | 0.22 |
| O1.3 | 7 | 11 | 0.19 | 0.21 |
| Z1.1 | 8 | 10 | 0.14 | 0.15 |
| Z1.2 | 9 | 8 | 0.13 | 0.14 |
| Z1.3 | 10 | 3 | 0.08 | 0.08 |
| Z1.4 | 11 | 4 | 0.08 | 0.08 |
| Z1.5 | 12 | 5 | 0.07 | 0.07 |

Table 6.1

Posterior probabilities of each radiocarbon determination being an outlier ($P_j = P[\phi_j = 1 \mid y]$), the 'Shroud of Turin', using a single prior probability $p_j = 0.1$. Complete sample, column (iv), and having removed determination O1.1 as a possible outlier, column (v).

From Table 6.1 we see that determination O1.1 ($j = 5$) has a noticeably large $P_j = P[\phi_j = 1 \mid y]$ of more than 0.3 and we might wish therefore to consider rejecting it as an outlier. This determination is the earliest (highest radiocarbon age) of all, see Figure 6.5. (Note that all determinations in Figure 6.5 can be identified using the score (j) shown in Table 6.1, column (ii), that gives the rank of each determination according to its radiocarbon age, eg. (5) = 12.)

It will always be difficult to decide when to remove determinations as outliers. In this case there is evidence, perhaps not conclusive, of determination

O1.1 being an outlier since $P_5 = 0.35$. However, determinations A1.1, O1.2 and O1.3 all have P_j 's that give us some evidence of these determinations being outliers. We decided only to remove the determination with the biggest P_j , that is O1.1. If we remove determination O1.1 and run the Gibbs sampler again with the remaining 11 determinations we obtain the posterior probabilities shown in Table 6.1, column (v). We see no further clear evidence of outliers this time. The corresponding distributions for θ can be seen in Figure 6.6. Also, for comparison, we have reproduced in the same Figure the original distribution using the complete sample.

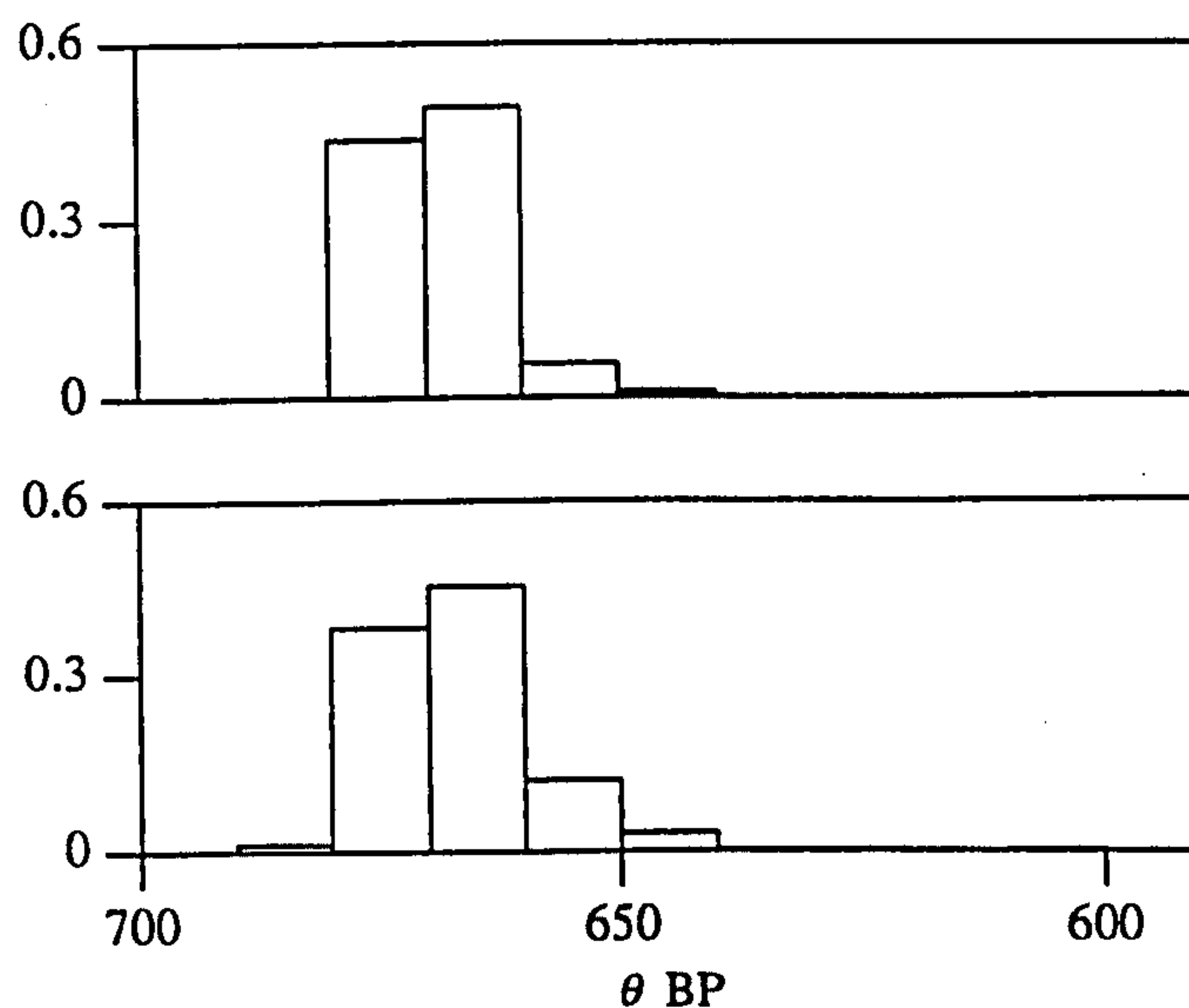


Figure 6.6
Histograms for the posterior distribution of θ , the 'Shroud of Turin',
with the original sample, found in Chapter 4, (above) and
with a possible outlier removed (below).

Note now, from Figure 6.5, that the range on the radiocarbon scale covered by the determinations 'compresses' to a much smaller range when transformed onto the calendar scale and for this reason the posterior distribution for θ is not wide spread. This is a result of the steep section of the calibration curve we are working on. Our analysis then gives only some evidence about discordancies in

the sample. This contrasts with the T -test, that does not include calibration, where the sample is considered 'not consistent'.

In this example the resulting posterior distributions, using the complete and the reduced samples, are nearly the same. This comes as a result of having, as part of our prior information, $\theta > 600$ which, in this case, gives robustness to our inferences. As stated before, the extent to which extreme determinations will affect our results in specific examples depends on the model and prior information used. The P_j 's above give us evidence about how in accordance the determinations are with

- (i) the model used,
- (ii) the prior information used
- and
- (iii) the rest of the determinations.

This accordance (or discordance) may or may not show in our results. However, the fact is that now we have a clear and explicit measure (P_j) for such 'accordance' that will provide us with evidence about how accurately the factors in our problem fit to each other. No comparable techniques have been developed within the context of radiocarbon dating and we have confidence that this represents an improvement to our general methodology that, hopefully, will result in more robust results. Below and in Chapter 7 we give further examples of this.

6.4.2 The summarisation problem: detecting outliers

We now turn to consider outliers in more complicated models where the θ_j 's are not necessarily equal, discussing how our methodology is applied in the 'summarisation problem'. We reconsider the summarisation problem studied in the introduction of the Chapter, originally considered by Pazdur and Krzanowski (1991). To illustrate our outlier methodology, we include in the sample the fake determination 1400 ± 70 .

From the discussion given Section 6.3, we see that the full conditionals and sampling techniques for β and α , the beginning and end of the *phenomenon* under study, remain the same as previously. We use the prior probability $p_j = 0.1$ for $j = 1, 2, \dots, 14$ of any determination being an outlier. We then calculate, using the Gibbs sampling technique, the posterior probability that each determination is an outlier, and these probabilities are reported in Table 6.2, columns (iii) and (vii).

| j | (j) | P_j | P_j | j | (j) | P_j | P_j |
|-----|-------|-------|-------|-----|-------|-------|-------|
| 1 | 4 | 0.10 | 0.09 | 8 | 7 | 0.09 | 0.09 |
| 2 | 2 | 0.10 | 0.10 | 9 | 5 | 0.13 | 0.13 |
| 3 | 3 | 0.10 | 0.09 | 10 | 13 | 0.09 | 0.17 |
| 4 | 10 | 0.10 | 0.10 | 11 | 12 | 0.09 | 0.10 |
| 5 | 9 | 0.09 | 0.09 | 12 | 6 | 0.10 | 0.09 |
| 6 | 11 | 0.11 | 0.11 | 13 | 8 | 0.08 | 0.08 |
| 7 | 1 | 0.09 | 0.09 | 14 | 14 | 0.39 | - |

Table 6.2

Posterior probabilities of each radiocarbon determination being an outlier ($P_j = P[\phi_j = 1 \mid y]$), Chancay culture summarisation problem, with $p_j = 0.1$.

Columns (iii) and (vii) include the fake determination 1400 ± 70 and columns (iv) and (viii) do not (original sample).

Note that $P_j = P[\phi_j = 1 \mid y]$ for the fake determination is noticeably large (0.39), whereas for the rest of the determinations it is approximately 0.1 (the same as the prior p_j 's) meaning that we have almost no new evidence of any of these

determinations being an outlier. However, the fake determination now is (*a posteriori*) nearly 4 times more likely to be an outlier on the basis of the current sample, thus prompting us to remove it. If we do so we obtain the posterior probabilities found in columns (iv) and (viii) of Table 6.2, with no further clear evidence of outliers. The posterior distributions of α and β should then be based on this corrected sample. This, of course, is the original sample.

Note that the probabilities P_j 's do not substantially change when calculated considering the fake determination or using the original sample (besides P_{14}). However, P_{10} does increase from 0.09 to 0.17. This determination is the earliest of the original sample. When P_{10} is calculated including the fake determination, y_{10} is well within the period for the Chancay culture, transformed onto the radiocarbon scale. If the fake determination is removed, y_{10} becomes the earliest and thus more likely to be an outlier, with respect to the current sample and prior information used (see Figure 6.2 in Section 6.1.3). Of course, the evidence that y_{10} is an outlier is only marginal. The posterior probability of each determination being an outlier does depend on the current sample, the model used and the priors $f(\theta \mid \psi)$ and $f(\psi)$.

On the other hand, Pazdur and Krzanowski (1991) in their analysis of the Chancay radiocarbon data, obtained the range 1000 to 470 BP (950 to 1480 AD) for the Chancay culture. This range coincides with our estimates using the original sample. However, to obtain their range, Pazdur and Krzanowski removed as 'outliers' the earliest and the latest of the determinations available. As stated above, using our methodology, we see no clear evidence of outliers in the sample. In fact, the latest of all determinations (y_7) has $P_7 = 0.09$. This means that now, *a posteriori*, that determination is less likely to be an outlier than before the sample was observed (since we started with the prior $p_7 = 0.1$). This contrasts and, perhaps, contradicts Pazdur and Krzanowski's heuristic procedure.

6.5 Remarks

We have developed, then, an extended general framework to tackle the problems of the analysis and interpretation of radiocarbon determinations. Now this extended framework allows for outliers. Note that, as opposed to other approaches, we have not developed a specific technique but a general methodology for the problem of outliers in radiocarbon data. The methodology can then be applied to a variety of dating problems using radiocarbon, by specifying the prior distributions $f(\theta | \psi)$ and $f(\psi)$. Note as well that the framework presented in Chapter 4 is a particular case of the framework presented here, if we let $p_j = 0$. That is, assigning a zero prior probability for outliers in the sample.

In presenting this methodology various alternatives come to our attention. We have used 'shifts' on the radiocarbon scale (δ_j 's) to explain outliers in radiocarbon data. We used the radiocarbon scale since outliers manifest *a posteriori* through the values of the determinations and therefore on the radiocarbon scale. An interesting modification would be to consider shifts on the calendar scale. For example, $y_j | \theta_j, \gamma_j, \phi_j \sim N(\mu(\theta_j + \gamma_j \phi_j), \sigma_j^2)$. This could model the possibility of a mistake in the assignment of a sample to a particular event or phase. Also we could consider using both shifts on the radiocarbon and calendar scales. However, we do not explore these possibilities any further.

We think that the techniques developed here can be of great benefit in the statistical investigation of a set of radiocarbon determinations. Also we are confident that, using this extended framework we can provide inferences, robust to the presence of outliers, making our assumptions explicit.

In the next Chapter we consider a dating problem arising from an archaeological site in Bavaria, Germany, where 13 radiocarbon determinations are available. For this example we have undertaken joint research with an archaeologist

and studied the excavation reports of the site. A complete model for the dating problem is suggested using the extended framework proposed in this Chapter. An outlier identification analysis is then developed within a complicated model including four archaeological phases and further prior information. It is then shown how the extended framework developed here can be easily applied to a real archaeological dating problem.

Chapter 7

The Galgenberg site: Robust estimates

7.1 Introduction

In this Chapter we concentrate our attention on how to model the characteristics of a particular archaeological dating problem and by using our techniques obtain robust estimates of the dates of the archaeological *phenomena* of interest. We will study a late Neolithic site in Germany and use the extended framework developed in the preceding Chapter to date a series of related events using the available radiocarbon determinations.

In previous Chapters, we have focussed our attention on developing our framework and on the technical details of how it is applied to a variety of archaeological examples. However little attention was given to the archaeology involved in these examples as we relied on other authors' considerations and assumptions. In contrast to this, now we concentrate on the archaeology involved in the problem by making our own interpretation of the site. We will study its characteristics and deduce a relative chronological sequence for some events observed within the stratigraphy. The archaeological dating problem will be defined and modelled mathematically. Then using our extended framework we will try to identify possible outliers within the determinations available. Finally we will obtain robust estimates for the dates of various events involved in the problem that will be interpreted within the archaeology of the site.

7.2 The mathematical modelling of a new dating problem

In this Section we briefly discuss the mathematical and statistical modelling of a new dating problem. This then will provide a basis for the development of the statistical model involved in the Galgenberg dating problem that we discuss in later Sections. In Chapter 4 we proposed a general modelling approach for the statistical analysis of sets of radiocarbon determinations. There we defined the vector y , denoting the radiocarbon determinations, θ , denoting the associated calendar years for the determinations (the calendar years in which the organic materials in the dated objects died) and ψ , which is taken to represent 'aspects' (calendar dates for events, lengths of phases etc.) of the archaeological dating *phenomena* of interest. The modelling process consists of,

- (i) defining $f(y \mid \theta)$ (the likelihood),
 - (ii) defining ψ according to the specific dating *phenomena* under study,
 - (iii) defining the prior distribution $f(\theta \mid \psi)$
- and
- (iv) defining the prior distribution $f(\psi)$.

In this Chapter we are interested on the 'elicitation' process necessary to complete the steps (i) to (iv) and so obtain a realistic and reliable mathematical model for the problem under study. To achieve this we will consider (i) to (iv) in turn.

To allow for outliers, in Chapter 6 we introduced the parameters δ_j and ϕ_j . The δ_j 's represent 'shifts', on the radiocarbon scale, for each determination and

the ϕ_j indicates whether determination y_j is an outlier or not (see Section 6.3). Given a set of radiocarbon determinations we have

$$y_j \mid \theta_j, \phi_j, \delta_j \sim N(\mu(\theta_j) + \phi_j \delta_j, \omega_j^2(\theta_j))$$

where $\omega_j^2(\theta_j) = \sigma_j^2 + \sigma^2(\theta_j)$, $\mu(\theta_j)$ is the piece-wise linear radiocarbon calibration curve and $\sigma^2(\theta_j)$ is the variance for the calibration curve (see Equations 1.3 and 3.14). It is then assumed that, given the θ_j 's, the y_j 's are independent and therefore we have

$$f(y \mid \theta, \delta, \phi) = \prod_{j=1}^m f(y_j \mid \theta_j, \phi_j, \delta_j).$$

$f(y \mid \theta, \phi_j, \delta_j)$ takes into account the calibration process, necessary for radiocarbon dating. Therefore our framework provides $f(y \mid \theta, \phi, \delta)$ and should not be considered part of the elicitation process.

Given the above, we see that the elicitation process, understood as the questions the statistician should ask in order to build a realistic statistical model for the dating problem under study, should then concentrate on (ii) to (iv) above. That is, defining ψ , $f(\theta \mid \psi)$ and $f(\psi)$. These distributions are stated *a priori* and thus are independent of (the value of) the determinations y . This implies that:

The mathematical modelling, as far as archaeologists are concerned, does not involve the values of the radiocarbon determinations. It does not involve either the radiocarbon calibration curve or the technicalities of the calibration process to be used.

We have stressed before that radiocarbon dating is a technique for dating objects and that dating in archaeology is, commonly, a more complex problem than that. In Chapter 4 we mentioned that a 'logical link' needs to be established between the dates for the objects and the archaeological *phenomena* of interest.

The dates for the objects, viewed in an isolated way, have little or no meaning. It is not until the relationship between such dates and the *phenomena* is understood that proper archaeological dating can be achieved.

In radiocarbon dating, the 'dates' for the objects dated refer, after calibration, to the calendar years in which the organic material contained in the objects died. We call these the associated calendar years for the objects, denoted by θ . Thus the distribution $f(\theta | \psi)$ models the relationship between the (associated calendar years for the) objects radiocarbon dated and the relevant 'aspects' of the archaeological dating *phenomena* ψ . This is the way that the 'logical link' between the objects dated and the *phenomena* is introduced into the analysis, by defining $f(\theta | \psi)$.

Therefore the elicitation process centres upon the following crucial question:

What is the relationship between the associated calendar years of the objects radiocarbon dated (θ) and the archaeological *phenomena* of interest?

Undoubtedly, this question will need to be addressed and properly answered before attempting to define $f(\theta | \psi)$ and $f(\psi)$. Indeed, before attempting to address the above question the archaeological dating *phenomena* must be defined and before trying to state $f(\theta | \psi)$ and $f(\psi)$, ψ must be defined. At this point the problem widens considerably and little can be said in a truly general sense about the choice for the distributions $f(\theta | \psi)$ and $f(\psi)$ and the definition of ψ . This can vary as much as types of dating problems do in archaeology. We have, however, presented in former Chapters a variety of examples (along with the example to be developed here) that give some ideas along these lines.

One issue that is worth considering relates to when in the elicitation process ψ should be defined. Certainly we need an appropriate definition for ψ to represent the relevant 'aspects' of the dating *phenomena*. It is clear that this cannot be achieved if the relationship between the associated calendar years and the *phenomena* has not been understood. Therefore, to conclude, we can say that the basis for the elicitation process involved in the modelling of a particular dating problem consists of the following sequence of steps.

- (i) To understand that at this point the values for the radiocarbon determinations are not relevant, nor are the technicalities of the calibration process.
- (ii) To address the question: What is it that we are trying to date, that is to say, what are the archaeological dating *phenomena* of interest?
- (iii) To address the question: What is the relationship between the associated calendar years for the objects radiocarbon dated and the archaeological dating *phenomena* of interest?
- (iv) In the light of (iii), try to define ψ .
- (v) Attempt to define $f(\theta \mid \psi)$ and $f(\psi)$ according to (iii) and (iv).

7.3 The Galgenberg site

Now we start our study of the Galgenberg dating problem. We first discuss the characteristics of the site and try to understand the dating problem at hand. In the next Section we will review other authors' approaches to the interpretation of the radiocarbon determinations available for the Galgenberg site. Second,

following the points (i) to (v) stated above, we will in Section 7.5 identify the archaeological dating *phenomena* for the Galgenberg and proceed to develop the mathematical model for interpreting the determinations.

The Galgenberg site is situated in the Isar valley, about 7km from Landshut, Bavaria. The site was first discovered in the early 1980's and has been excavated over a period of years. Although other occupation periods can be identified, the archaeological research has focussed on the 'Late Neolithic' activity. On the basis of the ceramics found, the Galgenberg has been classified as belonging to the 'Cham' Late Neolithic cultural group associated with that region. Broadly speaking, the time period expected for the Late Neolithic in that part of Bavaria is around 3000 to 2500 BC.

The Galgenberg is one of the few Cham group sites for which dating evidence is available. The site has a rôle to play in developing a chronology for the zone since it interacts with other Neolithic groups chronologically. By making the Galgenberg site chronology clearer it is expected that our understanding of the whole Neolithic chronology for the zone will be improved. In turn this will affect our understanding of other related cultural periods and of the archaeology of the area.

From our point of view, the Galgenberg presents a dating problem that has been studied over a number of years by other authors. As we will see, several questions have arisen in relation to the events we wish to date and those that can actually be dated, and the statistical analysis that should be followed. Thus the Galgenberg gives us an opportunity to show the potential of our framework and present what, we believe, is an acceptable and interesting analysis of this dating problem, in which we address some of the questions previously posed.

The archaeological research on the Galgenberg has been carried out by a British team and Ottaway (1988) presented a report on the excavations conducted up to that time. The Galgenberg is an oval site of around 40m by 60m in size and is surrounded by a ditch some 2m to 1.8m in depth. Several features have been studied. Stratigraphical relations, artefact typology and some faunal, molluscan, botanical and ceramic materials have been evaluated. Ottaway highlights two events within the occupation period of interest, both of which have been clearly identified, namely

(i) a fire of palisade fence posts,

and

(ii) a fire of posts near the entrance of the site.

These two fires can be identified in different parts of the stratigraphy and we will refer to them as the fires of fences (i) and (ii).

It is likely that the site was occupied at two different periods and the fires marked their (accidental or violent) end and can be taken to represent the termination of two different periods of occupation. As well, 'detailed artefact analysis' supports the presence of two occupation periods (Ottaway 1988).

Thirteen charcoal samples which can be related to the stratigraphy in the site were submitted for radiocarbon dating. The majority of the samples were taken from the burnt fences and the rest from other parts of the stratigraphy. Details of the corresponding radiocarbon determinations can be found in Table 7.1 (the way we group the determinations will be clarified in later Sections).

| Sample id. | Group I | Sample id. | Group III |
|------------|-----------|------------|-----------|
| GrN-12561 | 4255 ± 40 | GrN-12563 | 4150 ± 60 |
| GrN-14429 | 4310 ± 60 | GrN-12564 | 4210 ± 60 |
| | | UB-2551 | 4285 ± 85 |
| Sample id. | Group II | Sample id. | Group IV |
| GrN-12699 | 4510 ± 30 | GrN-12562 | 4290 ± 45 |
| GrN-12700 | 4225 ± 30 | | |
| GrN-12701 | 4280 ± 35 | | |
| GrN-12702 | 4385 ± 35 | | |
| GrN-14426 | 4420 ± 35 | | |
| GrN-14427 | 4245 ± 50 | | |
| GrN-14428 | 4500 ± 80 | | |

Table 7.1
Radiocarbon determinations available for the Galgenberg site,
taken from Aitchison *et al.* (1991).

7.3.1 Previous approaches

Ottaway (1988) used two approaches to analysing the radiocarbon determinations, though at that point in time she had only nine determinations available (determination UB-2551 and determinations GrN- x , for x less than 14426 in Table 7.1). Firstly she grouped those determinations arising from samples taken from the burnt materials from fences (i) and (ii) mentioned above. In total, she grouped seven of the nine determinations. She then had two groups, each one 'relating' to a fire. Using these, she calculated the pooled mean for each group. This gave her, after calibration, some evidence about the position in time for the fires. The calibration procedure used appears to be the 'calibrated confidence interval' technique explained in Section 2.3.4.

The pooled means for each group were 4350 ± 20 and 4180 ± 40 , corresponding to the materials of fence (i) and fence (ii). After calibrating both pooled means, Ottaway obtains the calendar time periods 3030-2920 BC and 2880-2700

BC, respectively. Ottaway (1988) explains that it is likely that these time periods

"(...) date two separate events, namely:

i) a fire of palisade fence posts between 3030-2920 Cal. BC, and ii) a fire of posts near the entrance between 2880 and 2700 Cal. BC."

However, the organic materials contained in the samples grouped cannot be considered to have died in the same calendar year. That is, the associated calendar years for the samples in each group cannot be assumed to be the same since they belong to different objects. Thus, strictly speaking, the determinations grouped cannot be considered as replicated samples and therefore the usage of pooled means for each group is not completely satisfactory (we have discussed this problem in Section 4.4.2).

Secondly, Ottaway (1988) considers the determinations as a whole and creates a 'dispersion diagram' on the radiocarbon scale to assess the occupation period for the site. This represents a graphical technique that does not include calibration and is intended to give a summary of the radiocarbon determinations. This statistical approach is clearly outlined in Ottaway (1986) where the problems of dispersion diagrams are analysed. Since no probabilistic model is used, it is not clear how to calibrate those dispersion diagrams. (We have reviewed both dispersion diagrams and this latter paper in Section 2.4.4.)

A further statistical analysis for the determinations arising from the Galgenberg was carried out in Aitchison *et al.* (1991). They develop a new statistical technique to overcome the problems of dispersion diagrams (in Section 2.4.4 we have reviewed the statistical techniques developed by Aitchison *et al.* 1991 giving our comments and critique). In this analysis the fires (i) and (ii) are not taken into account, nor any information about the stratigraphy in the site. They used all the thirteen determinations available and estimated the 'floruit' for the site. The

floruit is taken to represent

"(...) the period of time when the middle 50% of artefacts or materials from culture were produced (...)"

It is concluded that the estimated 95% interval for the floruit is (approximately) 3100 to 2800 BC. Since all the thirteen determinations were included in the analysis, such a time period estimates the floruit for the site as a whole.

The above approaches did not include in the analysis any information about the stratigraphy in the site nor the relationship between the associated calendar years for the charcoal samples and the fires of the fences (i) and (ii). We believe that a correct interpretation of the determinations available for the Galgenberg site can only be achieved if such crucial information is included in the statistical analysis. In the next Sections we will develop our own approach to the problem trying to include these considerations in the analysis. Then, at the end of Section 7.6, we will compare the above estimates with our results.

7.4 The dating problem

Two different events can be identified from the excavations at the Galgenberg site, namely, the fires of the fences (i) and (ii) mentioned before. These fires mark the ending of the two main occupation periods observed in the site. We then infer that the dating problem for the Galgenberg is to date those fires, expecting this to give us a clear chronology for the site.

The objects radiocarbon dated are charcoal samples, the majority belonging to the actual palisades or materials that formed the fences. The rest of the samples were taken from specific parts of the stratigraphy in the site and, broadly speaking, could be placed in relation to the fires (i) and (ii).

However, we must note that using radiocarbon dating, the fires of the fences cannot be directly dated. Some of the materials dated did belong to the fences but they cannot have died when the fires occurred. It is obvious that the construction of a fence must have occurred before the fire. Therefore, the organic materials forming the fences must have died before both the construction and the fire of the corresponding fence. The radiocarbon determinations then give us some direct evidence about the date for the construction of the fences only and thus we cannot aim at dating the fires (i) and (ii) directly. However, using the determinations, we could attempt to give early and, with further considerations, late boundaries for their position in time. That is, in archaeological terms, a 'terminus post quem' and, possibly, a 'terminus ante quem' for such events.

To simplify our discussion we identify the following events as

C1 - construction of fence (i),

F1 - fire of fence (i),

C2 - construction of fence (ii)

and

F2 - fire of fence (ii).

A key feature in the Galgenberg site is that C1, F1 and C2 and F2 are stratigraphically separated and must be chronologically ordered, the presence of two burnt horizons in several ditch sections appears to guarantee this sequence. That is, C1 and F1 must have occurred before C2 and F2. We then identify as the archaeological dating *phenomena* for this problem the events C1, F1, C2 and F2. From this, our approach to the dating problem will be as follows. Following the

guidelines for modelling discussed in Section 7.2, we will first deduce the relationship between the objects or materials radiocarbon dated and the relative chronology provided by C1, F1, C2 and F2 (the *phenomena*). Secondly, using the techniques presented in Chapters 4, we will be able to include such considerations in our modelling defining ψ , $f(\theta | \psi)$ and $f(\psi)$ accordingly and obtain a statistical analysis that will give us results interpretable in the light of the archaeological *phenomena* observed in the site, that is, C1, F1, C2 and F2.

7.4.1 Grouping the determinations

The radiocarbon samples are of two types, those belonging to the fences and those taken from the stratigraphy. From the former samples we create groups II and III and from the latter samples, groups I and IV. Now we give a brief description of the groups (the corresponding radiocarbon determinations are given in Table 7.1):

Group I, stratigraphically earliest:

- GrN-12561 Charcoal from western ditch butt, primary fill/deposit.
- GrN-14429 Charred stake in ditch bottom.

Group II, first destruction level in ditches (fence (i)):

- GrN-12699 Charred posts collapsed into western ditch.
- GrN-12700 "
- GrN-12701 "
- GrN-12702 "
- GrN-14426 Charred stake: '1. destruction level in ditch'.
- GrN-14427 "
- GrN-14428 "

Group III, second destruction level in ditches (fence (ii)):

GrN-12563 Charred collapsed palisade posts, eastern part of entrance of western ditch, destruction horizon above primary deposit.

GrN-12564 "

UB-2551 Charred collapsed palisade post, eastern ditch, apparently second destruction.

Group IV Latest stratigraphically:

GrN-12562 Charcoal from final recut of western forework.

What is now necessary is to understand the relationships between these groups of samples and the archaeological *phenomena* observed in the site, in other words, to understand the chronological relationships of the above groups with the events C1, F1, C2 and F2. By understanding the logical sequence of events we will suggest such a relationship, thus finding a 'logical link' between the objects radiocarbon dated and the *phenomena* C1, F1, C2, F2.

7.4.2 The relative chronology

As pointed out before, the wood used to construct the fences must have died before the corresponding fence was erected and, subsequently, burnt. Thus the associated calendar years for the determinations included in groups II and III must be earlier than C1 and C2, respectively.

Now, it is very likely that the ditches found at the Galgenberg site were dug while the construction of the fences was in progress. In fact, it is probable that the non-wooden materials forming the ramparts (turf, soil) were dug out from the ditches. Assuming that the deposition of material in group I is indeed the earliest fill in the ditch then the date of deposition must be close in time to the construction of fence (i), that is, C1. The organic materials contained in the samples

forming group I should have died before deposition and thus we estimate that their associated calendar years should be, at the latest, contemporaneous with C1. In other words, the associated calendar years for the determinations included in group I must be earlier than or contemporaneous with C1.

Furthermore, there is evidence that the sample in group IV is the latest stratigraphically. Assuming that this sample is indeed a product of the late activity in the site, we can say that its associated calendar year postdates F2.

Thus we have now 'located' all the samples within the relative chronology given by C1, F1, C2 and F2. To clarify this discussion, in Figure 7.1 we present a diagram representing the chronological ordering of the groups of determinations I, II, III, and IV with such events.

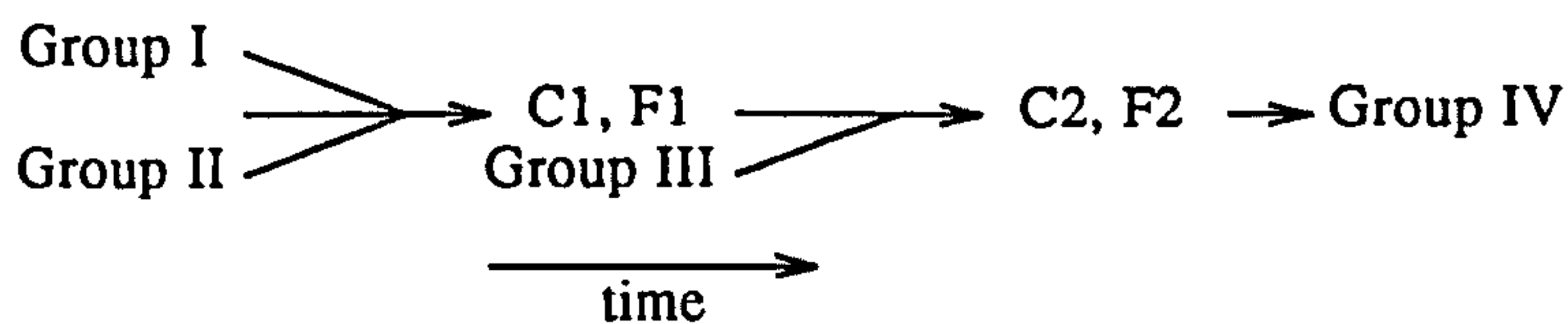


Figure 7.1

Chronological relationships assumed for the Galgenberg site.

Continuing with our analysis of the objects radiocarbon dated we can add the following considerations to the relative chronology stated above. It is very likely that the wood used in the site, for building fences or other purposes, belonged to relatively short-lived trees. This arises as a technological constraint since small short-lived trees are easier to fell and cut. Assuming this is the case, and no wood was reused, it can be expected that the internal rings in the trees had a maximum age of 50 to 70 years at felling (see Barfield 1991, figure 1, for a convenient analogy). However, we adopt a more conservative duration of 100 years for the wood used in the site.

Furthermore, since the associated calendar years for the determinations of both group I and II predate C1, the latest *possible* age for the tree-rings in the samples of those groups should be the same. Similarly, the latest *possible* age for the rings in the samples of group III should be before C2. To clarify this we represent the following events by:

α_1 - earliest possible age for the rings in the samples of group I.

α_2 - earliest possible age for the rings in the samples of group II.

β - latest possible age for the rings in the samples of groups I and II.

α_3 - earliest possible age for the rings in the samples of group III.

β_3 - latest possible age for the rings in the samples of group III.

and

$\theta_{4,1}$ - associated calendar year for the sample in group IV.

From this we see that the above discussion on the age of the wood used in the site means that (using years BP)

$$0 < \alpha_1 - \beta < 100, \quad 0 < \alpha_2 - \beta < 100 \quad \text{and} \quad 0 < \alpha_3 - \beta_3 < 100.$$

At this point we are now able to give a complete description of the assumed relative chronology for the Galgenberg site, combining α_1 , α_2 , β , α_3 , β_3 and $\theta_{4,1}$ with the events C1, F1, C2 and F2. We do this using the diagram presented in Figure 7.2.

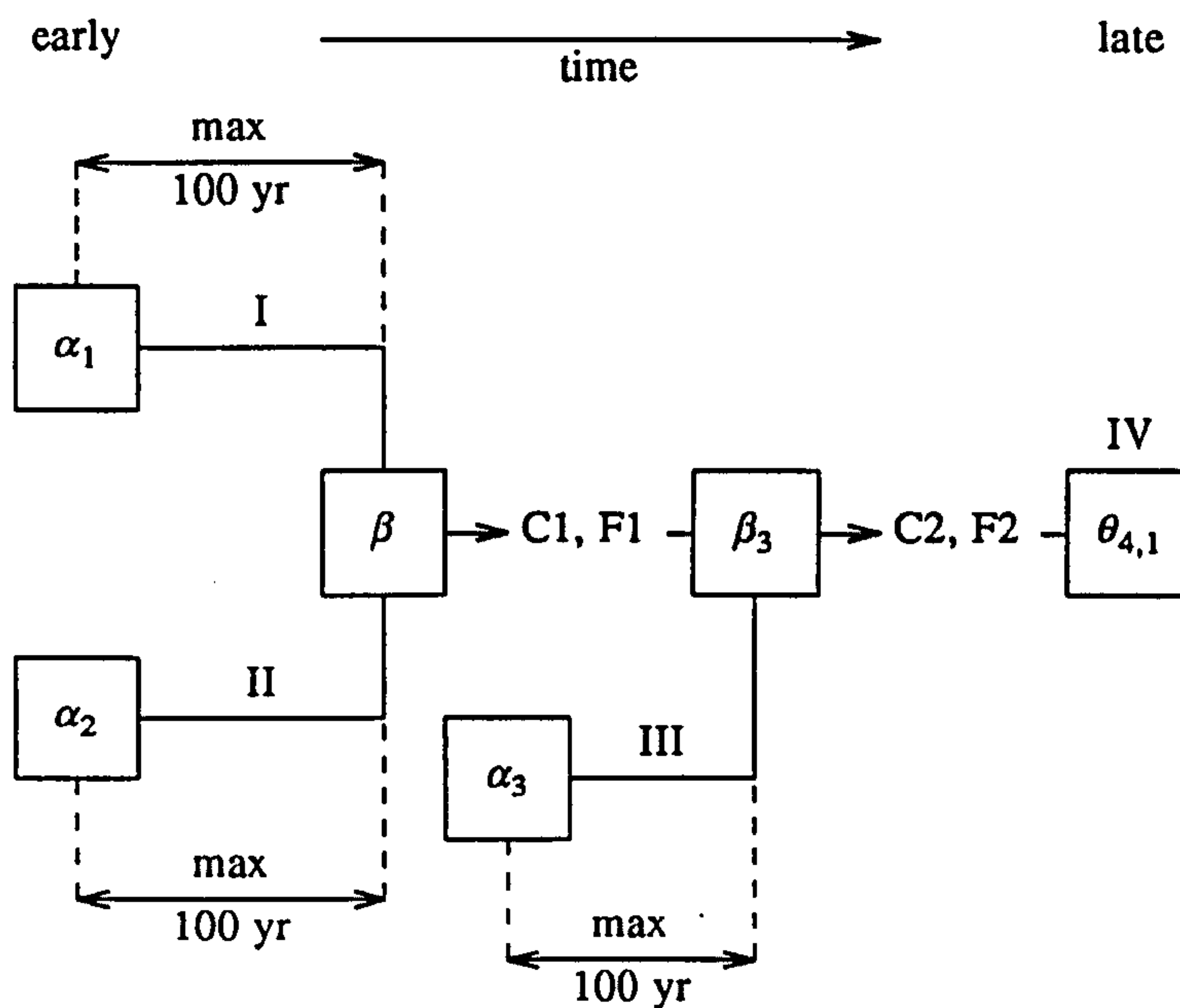


Figure 7.2

Relative chronology for the Galgenberg site.

Roman numbers represent where the corresponding groups of determinations should 'lie'.

In the next Section we will build the complete mathematical model for the Galgenberg dating problem. Note that, as stated before, we have not considered the values of the radiocarbon determinations to build this relative chronology and these will not be used in the model. The determinations will be used later in Section 7.5 to find the posterior distributions of interest.

7.4.3 Mathematical modelling

In this Section we develop the formal mathematical model for the Galgenberg dating problem. We will model the chronological relationships between the different events defined in the site using the techniques discussed in Chapters 4 and 5. We will also adopt the robustness techniques presented in Chapter 6, and

in the following Sections present an outlier identification analysis and our final estimates.

Using the notation and techniques presented in Section 5.6, it is not difficult to obtain a model for the relative chronology assumed in the Galgenberg dating problem. Let $y_{i,j} \pm \sigma_{i,j}$ be the j^{th} determination from group i associated with the calendar year $\theta_{i,j}$. As explained before, in Section 7.2, we need to define ψ and the prior distributions $f(\theta \mid \psi)$ and $f(\psi)$. This with respect to the dating problem of interest and the relationship between the determinations and the dating *phenomena*. We will proceed as follows.

The associated calendar years $\theta_{i,j}$ can all be located within the relative chronology for the site explained above. For example, given α_1 and β we know that $\theta_{1,j} \in (\beta, \alpha_1)$ with probability one (see Figure 7.2). To facilitate our discussion we state that, given ψ , each $\theta_{i,j}$ belongs to the phase ψ_{2i} to ψ_{2i-1} . Then we define

$$\psi = (\psi_1, \psi_2, \dots, \psi_8) = (\beta, \alpha_1, \beta, \alpha_2, \beta_3, \alpha_3, 0, \beta_3)$$

and we assume that

$$\theta_{i,j} \mid \psi \sim U(\psi_{2i-1}, \psi_{2i})$$

and

$$f(\theta \mid \psi) = \prod_{i=1}^4 \prod_{j=1}^{m_i} f(\theta_{i,j} \mid \psi).$$

(Here we have $m_1 = 2$, $m_2 = 7$, $m_3 = 3$ and $m_4 = 1$.)

That is, we translate the early and late boundaries for the phase containing each group to $\alpha_1, \alpha_2, \beta, \alpha_3, \beta_3$. For group IV, β_3 provides an early boundary, but no late boundary is defined within the relative chronology of the site. Thus we have defined $\psi_7 = 0$. Furthermore, since we have no prior information about the position of each $\theta_{i,j}$ within the phase ψ_{2i} to ψ_{2i-1} , we use a uniform prior

distribution for $\theta_{i,j}$ within that phase.

The relative chronology is completely modelled when we define $f(\psi)$ as the distribution with support in

$$\beta > \beta_3,$$

and

$$0 < \alpha_1 - \beta < 100, \quad 0 < \alpha_2 - \beta < 100 \quad \text{and} \quad 0 < \alpha_3 - \beta_3 < 100.$$

To clarify the ideas we present, in Figure 7.3, the corresponding hierarchy diagram for the modelling of the Galgenberg dating problem (following the conventions given in Section 5.6.2).

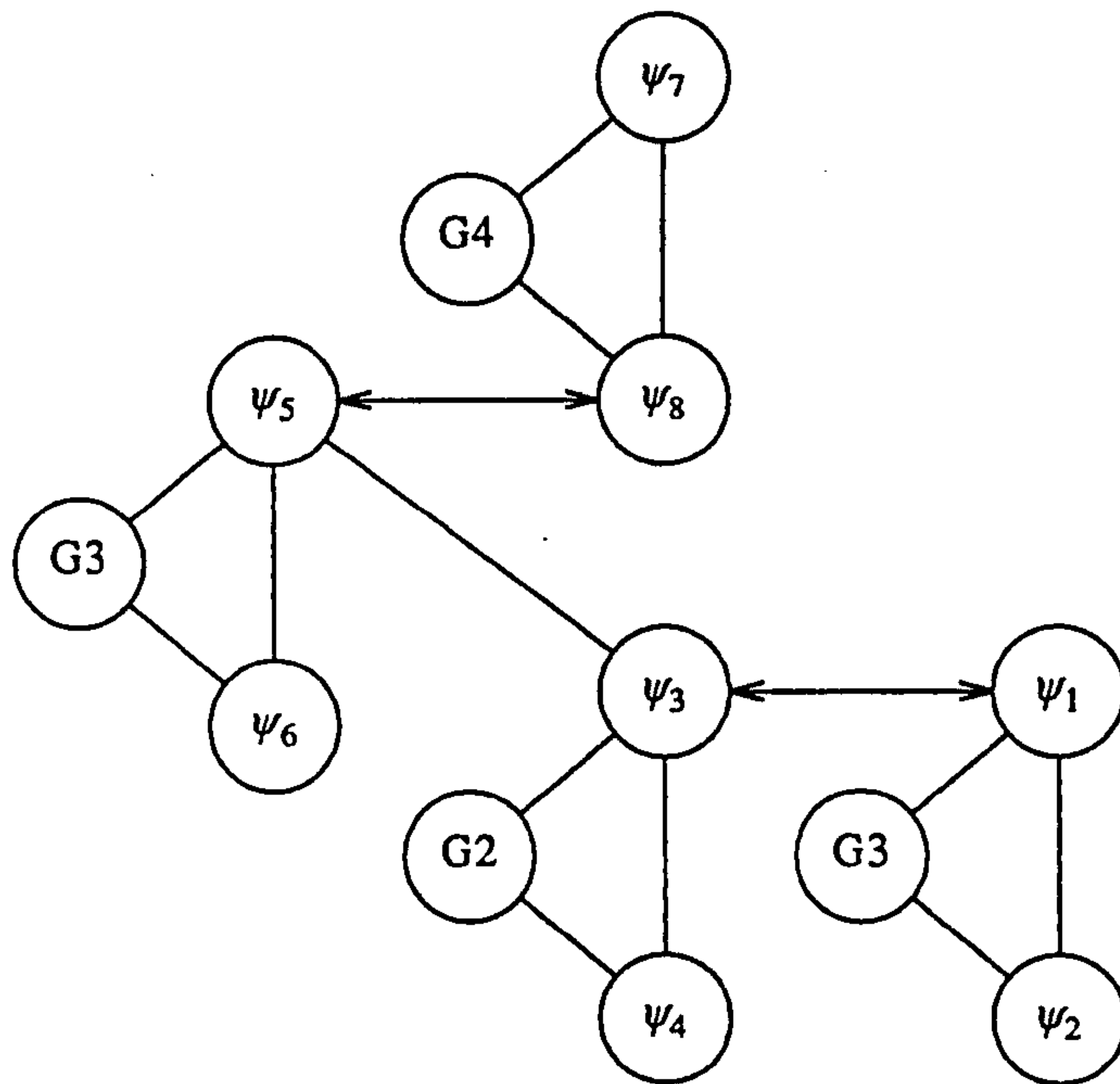


Figure 7.3

Hierarchy diagram for the Galgenberg modelling.

We identify $\psi = (\beta, \alpha_1, \beta, \alpha_2, \beta_3, \alpha_3, 0, \beta_3)$.

We will include the robustness techniques discussed in Chapter 6, using the model

$$y_{i,j} \mid \theta_{i,j}, \delta_{i,j}, \phi_{i,j} \sim N(\mu(\theta_{i,j}) + \phi_{i,j} \delta_{i,j}, \omega_{i,j}^2(\theta_{i,j}))$$

(where $\omega_{i,j}^2(\theta_{i,j}) = \sigma_{i,j}^2 + \sigma^2(\theta_{i,j})$). In this case, and only for practical reasons, we do not consider the variance in the calibration curve $\sigma^2(\theta)$ since the determinations used are not high-precision (the standard deviations are higher than 30 years). That is, we assume $\omega_{i,j}^2(\theta_{i,j}) \approx \sigma_{i,j}^2$. Thus it is not difficult to calculate the full conditionals for all the parameters of interest using the techniques presented in Section 5.6 and 6.4. These are

$$f(\theta_{i,j} \mid y, \theta_{-(i,j)}, \psi, \delta, \phi) = f(\theta_{i,j} \mid y_{i,j}, \psi_{2i-1}, \psi_{2i}, \delta, \phi) \propto \exp\left\{-\frac{1}{2}\left(\frac{(\mu(\theta_{i,j}) - (y_{i,j} - \phi_{i,j} \delta_{i,j}))^2}{\sigma_{i,j}^2}\right)\right\} I(\theta_{i,j})_{(\psi_{2i-1}, \psi_{2i})},$$

for $i = 1, 2, 3, 4$ and $j = 1, 2, \dots, m_i$. Since

$$f(\psi_i \mid y, \theta, \psi_{-i}, \delta, \phi) = f(\psi_i \mid \theta, \psi_{-i}) \propto f(\theta \mid \psi) f(\psi_i \mid \psi_{-i})$$

for $i = 1, 2, 3, 4$ we have

$$f(\alpha_1 \mid \theta, \alpha_2, \beta, \alpha_3, \beta_3) \propto (\alpha_1 - \beta)^{-2} I(\alpha_1)_{(\max(\theta_1), \beta + 100)},$$

$$f(\alpha_2 \mid \theta, \alpha_1, \beta, \alpha_3, \beta_3) \propto (\alpha_2 - \beta)^{-7} I(\alpha_2)_{(\max(\theta_2), \beta + 100)},$$

$$f(\beta \mid \theta, \alpha_1, \alpha_2, \alpha_3, \beta_3) \propto (\alpha_1 - \beta)^{-2} (\alpha_2 - \beta)^{-7} I(\beta)_{(a, b)}$$

where

$$a = \max(\max(\alpha_1, \alpha_2) - 100, \beta_3) \text{ and}$$

$$b = \min(\min(\theta_1), \min(\theta_2)),$$

and

$$f(\alpha_3 \mid \theta, \alpha_1, \alpha_2, \beta, \beta_3) \propto (\alpha_3 - \beta_3)^{-3} I(\alpha_3)_{(\max(\theta_3), \beta_3 + 100)}$$

and

$$f(\beta_3 \mid \theta, \alpha_1, \alpha_2, \beta, \alpha_3) \propto (\alpha_3 - \beta_3)^{-3} \beta_3^{-1} I(\beta_3)_{(\max(\alpha_3 - 100, \theta_{4,1}), \min(\theta_3))}.$$

The remainder of the full conditionals are

$$\delta_{i,j} \mid y, \theta, \delta_{-(i,j)}, \phi \sim N(y_{i,j} - \mu(\theta_{i,j}), \sigma_{i,j}^2), \text{ or } \phi_{i,j} = 1,$$

and

$$f(\phi_{i,j} \mid y, \theta, \delta, \phi_{-(i,j)}) \propto p_{i,j}^{\phi_{i,j}} (1 - p_{i,j})^{1 - \phi_{i,j}} \exp \left\{ -\frac{1}{2} \left(\frac{(\mu(\theta_{i,j}) - (y_{i,j} - \delta_{i,j} \phi_{i,j}))^2}{\sigma_{i,j}^2} \right) \right\},$$

where $p_{i,j}$ is the prior probability that determination $y_{i,j} \pm \sigma_{i,j}$ is an outlier.

In Section 5.6 we have discussed general guidelines for sampling from the above full conditionals. Once we have appropriate sampling routines, a MCMC method can be implemented to find the posterior distributions required. In particular the posterior probabilities $P_{i,j} = P[\phi_{i,j} = 1 \mid y]$, will give us evidence about the presence of outliers in the sample. We present and discuss these results in the next Sections.

7.5 Outlier identification

Given the model developed above, we are now able to combine consistently the considerations about the relative chronology of events in the Galgenberg site and the radiocarbon determinations available. Our initial step will be to analyse the posterior probabilities

$$P_{i,j} = P[\phi_{i,j} = 1 \mid y]$$

to identify possible outliers. As discussed in Chapter 6, these probabilities provide us with evidence about how well in accordance determination $y_{i,j}$ (the j^{th} determination from group i) is with,

- (i) the rest of the determinations,
- and
- (ii) the relative chronology for the site.

In Table 7.2 we present each of these probabilities for the thirteen determinations involved in the problem. As suggested in Chapter 6, we are using a single prior probability of 0.1 for any determination being an outlier. That is, *a priori*, we expect 1 in 10 determinations to be discordant or outliers. In Figure 7.4 we present the posterior marginal distributions for the parameters $\alpha_1, \alpha_2, \beta, \alpha_3, \beta_3$ and $\theta_{4,1}$ of the Galgenberg chronology.

| Sample id. | $P_{1,j}$ | Sample id. | $P_{3,j}$ |
|------------|-----------|------------|-----------|
| GrN-12561 | 0.07 | GrN-12563 | 0.08 |
| GrN-14429 | 0.09 | GrN-12564 | 0.08 |
| | | UB-2551 | 0.14 |
| Sample id. | $P_{2,j}$ | Sample id. | $P_{4,1}$ |
| GrN-12699 | 0.98 | GrN-12562 | 0.42 |
| GrN-12700 | 0.10 | | |
| GrN-12701 | 0.07 | | |
| GrN-12702 | 0.07 | | |
| GrN-14426 | 0.10 | | |
| GrN-14427 | 0.08 | | |
| GrN-14428 | 0.18 | | |

Table 7.2
Posterior probabilities of each determination being an outlier
for the Galgenberg dating problem.

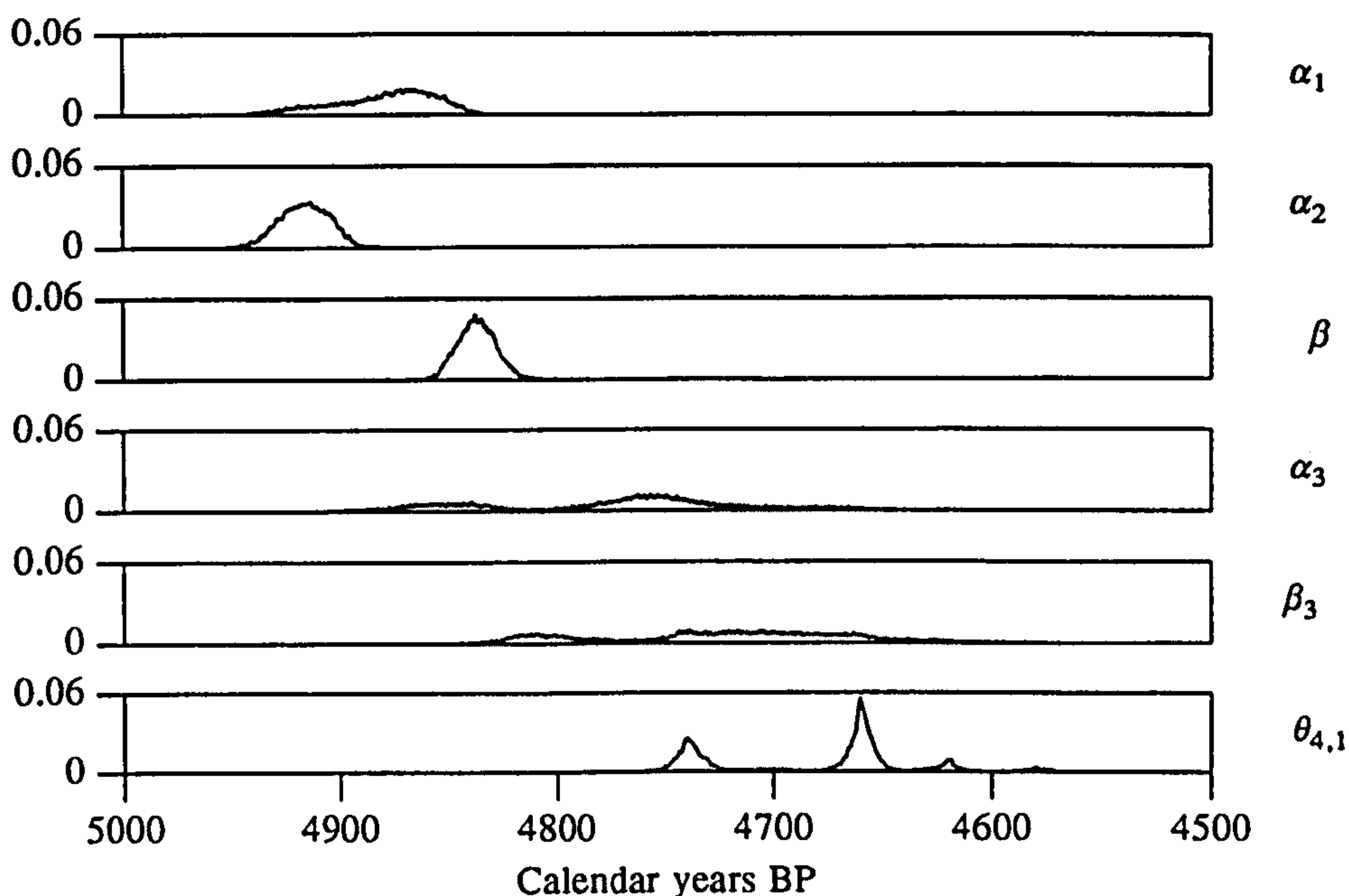


Figure 7.4
Marginal posterior distributions for the parameters of
the chronology of the Galgenberg dating problem.

Apart from two posterior probabilities the rest of the $P_{i,j}$'s shown in Table 7.2 remain basically the same as the prior value of 0.1. That is, we have no conclusive evidence, *a posteriori*, that the corresponding determinations are misleading (outliers) given that, *a priori*, we expected 1 in 10 determinations to be outliers. The two determinations that do have conspicuous posterior probabilities $P_{i,j}$'s are GrN-12699, from group II (with 0.98), and GrN-12562, from group IV (with 0.42).

$P_{2,1}$ attributed to determination GrN-12699 is very large (> 0.9). This determination differs from the other determinations in its group, having the highest radiocarbon age of all and a very low standard deviation (30 years, see Table 7.1). The rest of the determinations in its group have more or less the same $P_{i,j}$ as the prior, 0.1, and therefore we have no new evidence about those determinations being outliers. Thus we have very conclusive evidence that determination

GrN-12699 is not in accordance with the rest of the determinations and the prior information about the relative chronology of the site. It seems clear that, to obtain robust and reliable estimates, we should remove this determination and re-run our analysis using the corrected set of determinations (we will do this in the next Section).

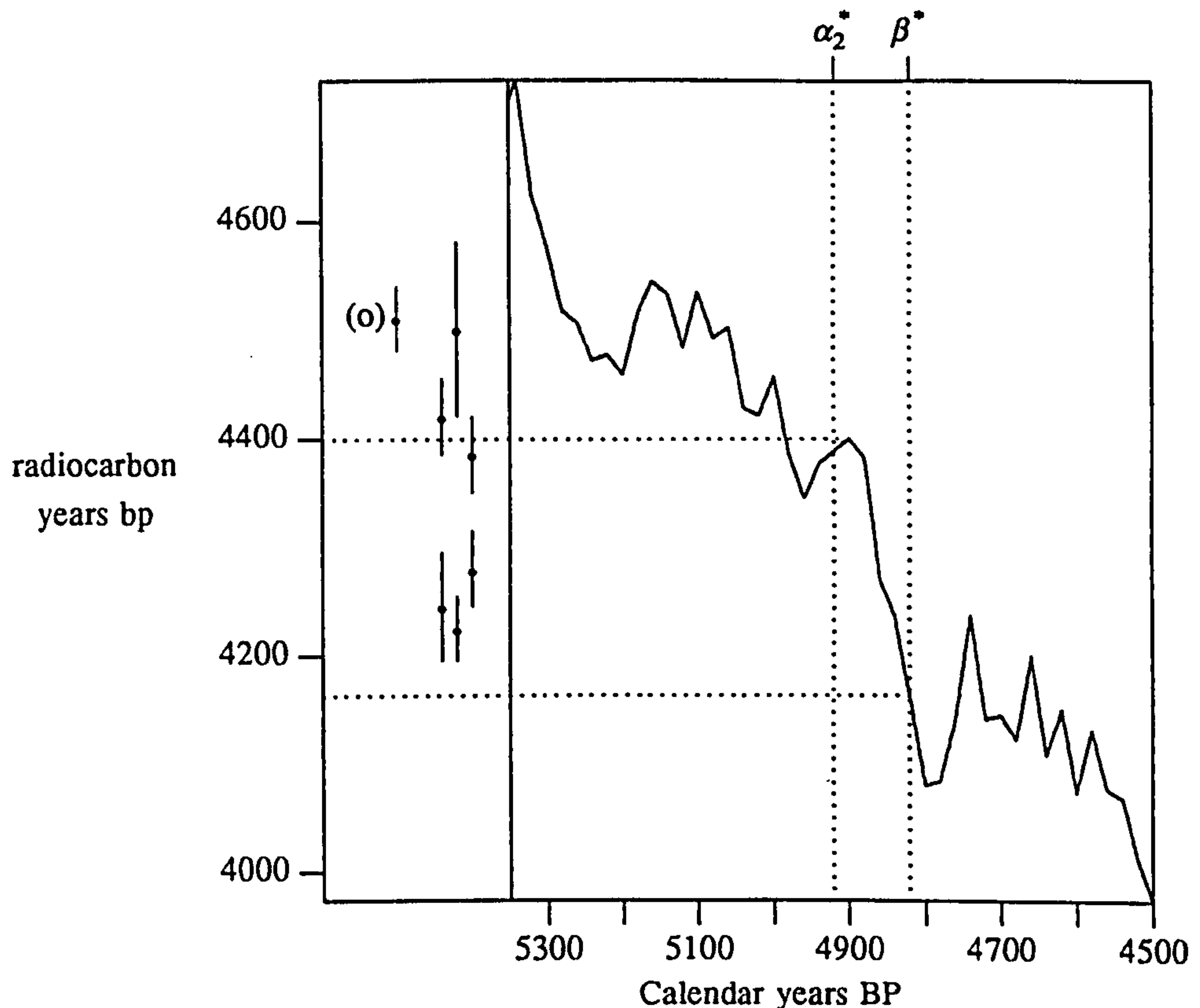


Figure 7.5

Section of the piece-wise linear calibration curve along with the radiocarbon determinations from group II and the range $(\mu(\beta^*, \alpha_2^*))$. (o) corresponds to determination GrN-12699 detected as a probable outlier.

To illustrate the above we proceed as follows. We note that the $\theta_{i,j}$'s of group II are contained in (β, α_2) . Since the posterior distributions of α_2 and β are more or less symmetric and not wide-spread, the MAP estimators, α_2^* and β^* , give good summaries for the posterior distributions of these parameters. Therefore, we would expect the determinations of group II to lie approximately within

the range $\mu(\beta^*, \alpha_2^*)$. Considering this, we present in Figure 7.5 the determinations of group II, together with α_2^* , β^* , $\mu(\beta^*, \alpha_2^*)$ and the relevant part of the calibration curve. It can then be seen that determination GrN-12699 is about 4 to 5 standard deviations away from the range $\mu(\beta^*, \alpha_2^*)$, as opposed to the rest of the determinations in its group, which are within approximately one standard deviation of that range.

A more complicated case is the one presented by determination GrN-12562. The corresponding $P_{4,1}$ is more than 0.4. This gives us some, although perhaps not conclusive, evidence that the determination is not in accordance with the rest of the sample and the prior information used in the analysis. Therefore it might be considered as an outlier. We must note, however, that this determination is the only one in group IV. It might seem paradoxical that a determination that stands alone in a group has a strong probability of being an outlier.

How can we call a determination that stands alone an outlier?

This apparent paradox is easily explained by noticing that determination GrN-12562 does *not* stand alone, simply because it is *not* interpreted on its own. There is detailed prior information that links this determination with the rest. Specifically, we stated that it was the 'latest stratigraphically' ($\beta_3 > \theta_{4,1}$). Therefore, we can still say that the determination is, to some extent, not in accordance with the other determinations (in other groups) and with the prior information used (the relative chronology of the site). To illustrate this we present, in Figure 7.6, the 95% highest posterior density region for β_3 transformed onto the radiocarbon scale, along with determination GrN-12562 and the relevant part of the calibration curve. We see then that determination GrN-12562 is 1 to 5 standard deviations above that range. According to our model, we would have expected it to be within or below that range.

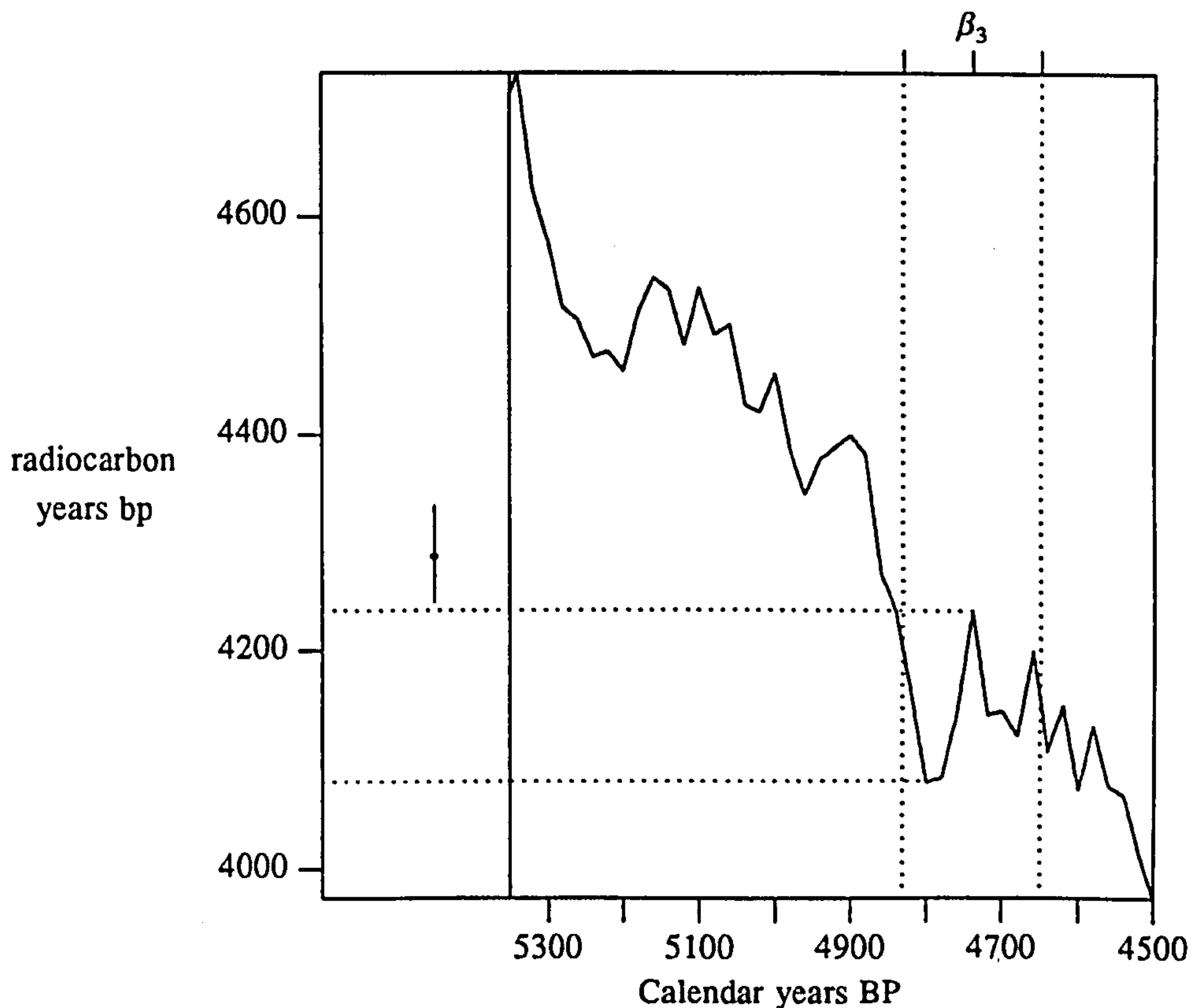


Figure 7.6
Section of the piece-wise linear calibration curve
along with radiocarbon determination GrN-12562 (group IV) and the
95% HPD region for β_3 transformed onto the radiocarbon scale.
Determination GrN-12562 is identified as a possible outlier.

However, since there are no other determinations in group IV that might help us balance what is happening, in this case it is difficult to decide whether determination GrN-12562 is erroneous (an outlier) or whether there is a piece of our prior information that is not accurate (related to the assumption that $\beta_3 > \theta_{4,1}$). It seems, though, that if we wish to make robust inferences and derive robust estimates from these determinations, then determination GrN-12562 should not be considered.

7.6 Robust estimates

We now reconsider the Galgenberg dating problem but remove determinations GrN-12699, from group II, and GrN-12562, from group IV, which we consider to be outliers. Removing determination GrN-12699 represents no technical difficulty since we only need to reduce the number of elements in group II to 6, that is $m_2 = 6$. However, removing determination GrN-12562 does present some complications since this determination was the only one forming group IV and now we do not have any determination associated with the 'late activity' in the site. Thus, our relative chronology changes slightly since now we do not have any 'terminus ante quem' for the fire of fence (ii) (or F2). This revised relative chronology is outlined in Figure 7.7 where $\theta_{4,1}$ is removed.

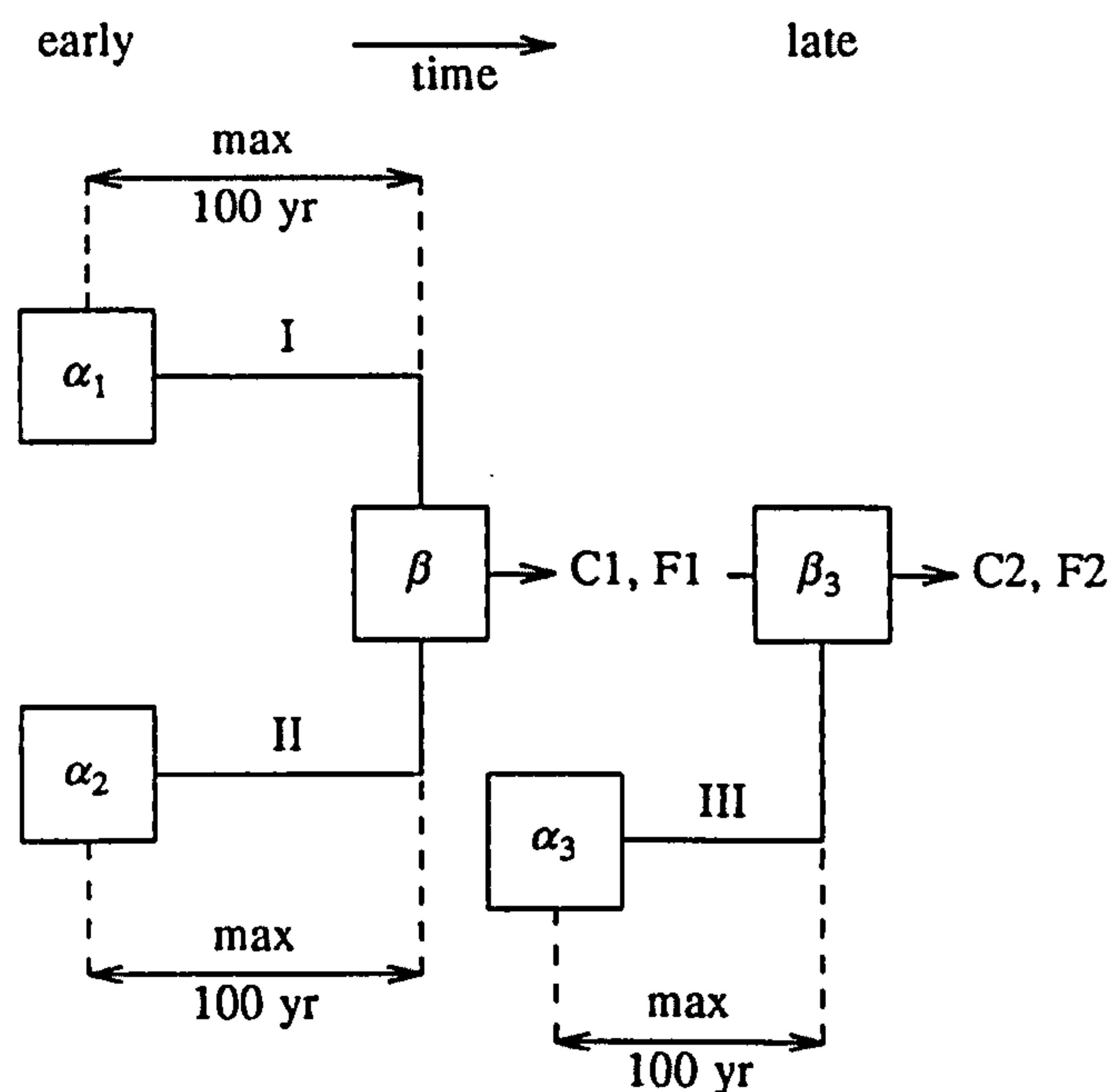


Figure 7.7

Relative chronology for the Galgenberg site, with $\theta_{4,1}$ not included in the sequence.

Roman numbers represent where the corresponding groups of determinations should 'lie'.

The mathematical modelling needs slight alteration, but making the necessary modifications presents no technical difficulty. We avoid going through further technical details and concentrate on interpreting the results.

We re-ran our analysis omitting determinations GrN-12699 and GrN-12562 and obtained the posterior probabilities of each determination being an outlier shown in Table 7.3. The corresponding marginal distributions for the parameters are shown in Figure 7.8.

| Sample id. | $P_{1,j}$ | Sample id. | $P_{3,j}$ |
|------------|-----------|------------|-----------|
| GrN-12561 | 0.07 | GrN-12563 | 0.08 |
| GrN-14429 | 0.09 | GrN-12564 | 0.11 |
| | | UB-2551 | 0.18 |
| Sample id. | $P_{2,j}$ | Sample id. | $P_{4,1}$ |
| GrN-12699 | rmv. | GrN-12562 | rmv. |
| GrN-12700 | 0.10 | | |
| GrN-12701 | 0.07 | | |
| GrN-12702 | 0.07 | | |
| GrN-14426 | 0.10 | | |
| GrN-14427 | 0.08 | | |
| GrN-14428 | 0.19 | | |

Table 7.3

Posterior probabilities of each determination being an outlier.
The Galgenberg dating problem, with determinations
GrN-12699 and GrN-12562 removed as possible outliers.

We consider the $P_{i,j}$'s shown in Table 7.3. The posterior probabilities observed remain at around the prior value of 0.1 and thus we have, *a posteriori*, no further clear evidence for outliers in the sample. The resulting posterior distributions do not vary greatly from the marginal distributions obtained with the former analysis, as seen in Figures 7.4 and 7.8. However, since in this case we have removed $\theta_{4,1}$, the earliest date in the model, the distributions of α_3 and β_3 adjust to this. They are now less spread and shifted about 100 years later.

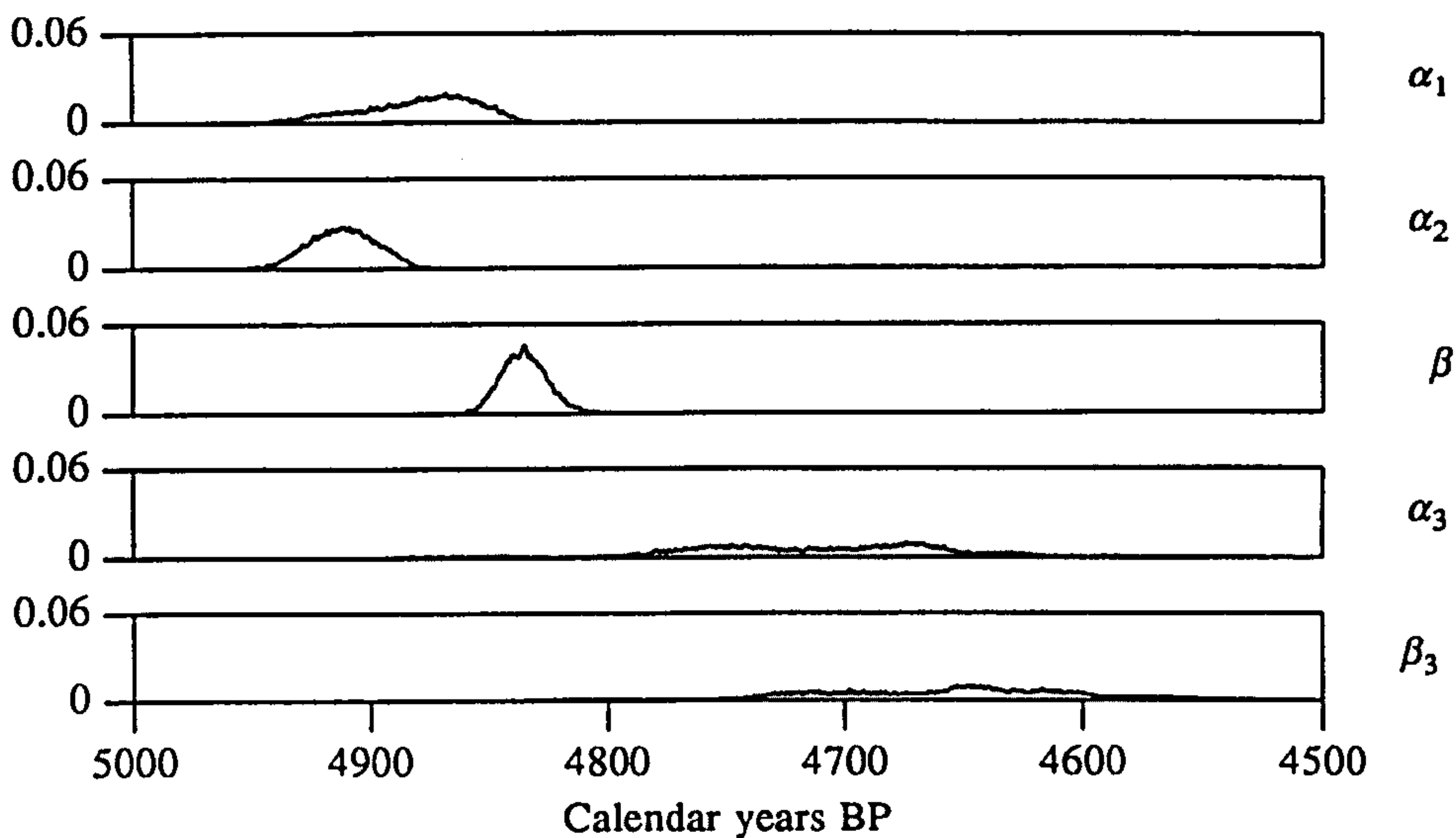


Figure 7.8
Posterior marginal distributions for the relevant parameters
of the Galgenberg dating problem, with two determinations
removed as possible outliers.

At this point we are able to interpret the posterior distributions of interest and give our final estimates. The interpretation of the results is focussed on the relevant archaeological *phenomena* observed in the site. That is, the construction and fire of the fences (i) and (ii), or C1, F1, C2 and F2. As seen in the diagram presented in Figure 7.7, the sequence of events C1, F1 is predated by β and postdated by β_3 . Thus we have a 'terminus post quem' and a 'terminus ante quem' for these events. The sequence C2, F2 is predated by β_3 but we have no parameter that postdates it. Thus, in this case, we only have a 'terminus post quem' for the events.

We then concentrate our interpretation on the posterior distributions $\beta | y$ and $\beta_3 | y$. In principle we should try to interpret the joint posterior distribution $\beta, \beta_3 | y$. Fortunately for us these two parameters seem, *a posteriori*, to be independent since their correlation is estimated to be -0.007 (estimated from the MCMC sample obtained) and thus we do not lose information by considering the

marginals only. As pointed out above, the sequence C1, F1 should have occurred after β and before β_3 and the sequence C2, F2 after β_3 . The posterior distributions $f(\beta | y)$ and $f(\beta_3 | y)$ then represent the current knowledge, based on the relative chronology of the site and the radiocarbon determinations available, about the position in time (on the calendar scale) of β and β_3 . What remains now is to examine those posterior distributions and observe where in time C1, F1, C2 and F2 should be located. The 95% highest posterior density regions of these distributions are (approximately) 4850 to 4820 BP for β and 4740 to 4560 BP for β_3 . Broadly speaking, we can say that the sequence C1, F1 lies in the region 4850 to 4820 BP and 4740 to 4560 BP (and C2 after that last range). However, given the relative chronology of the site, we believe that in this case a better interpretation of the distributions $f(\beta | y)$ and $f(\beta_3 | y)$ can be achieved if we analyse the corresponding cumulative distribution functions.

Since the sequence C1, F1 should have occurred after β , then C1, F1 occurred after any time t_1 with probability $P[t_1 < \beta | y]$. Similarly, F1 occurred before any time t_2 with probability $P[t_2 > \beta_3 | \theta]$ and C2 occurred after any time t_3 with probability $P[t_3 < \beta_3 | y]$. These probabilities are or can be calculated from the cumulative distribution functions of β and β_3 . In Figure 7.9 we present these distributions together with the probable position in the calendar scale for the events C1, F1, and C2.

The position in time for the events C1, F1 and C2 provided in Figure 7.9 are only tentative and must be interpreted as time ranges (in the direction of the arrows attached to each event) rather than point estimates. Note that, since we do not have a 'terminus post quem' for C2 (in this corrected model), its time range is not bounded. Using the distributions presented in Figure 7.9, any probabilities at any level or inter-quantile ranges can be approximated for β and β_3 . From that, probable time ranges for C1, F1 and C2 can be estimated with any desired

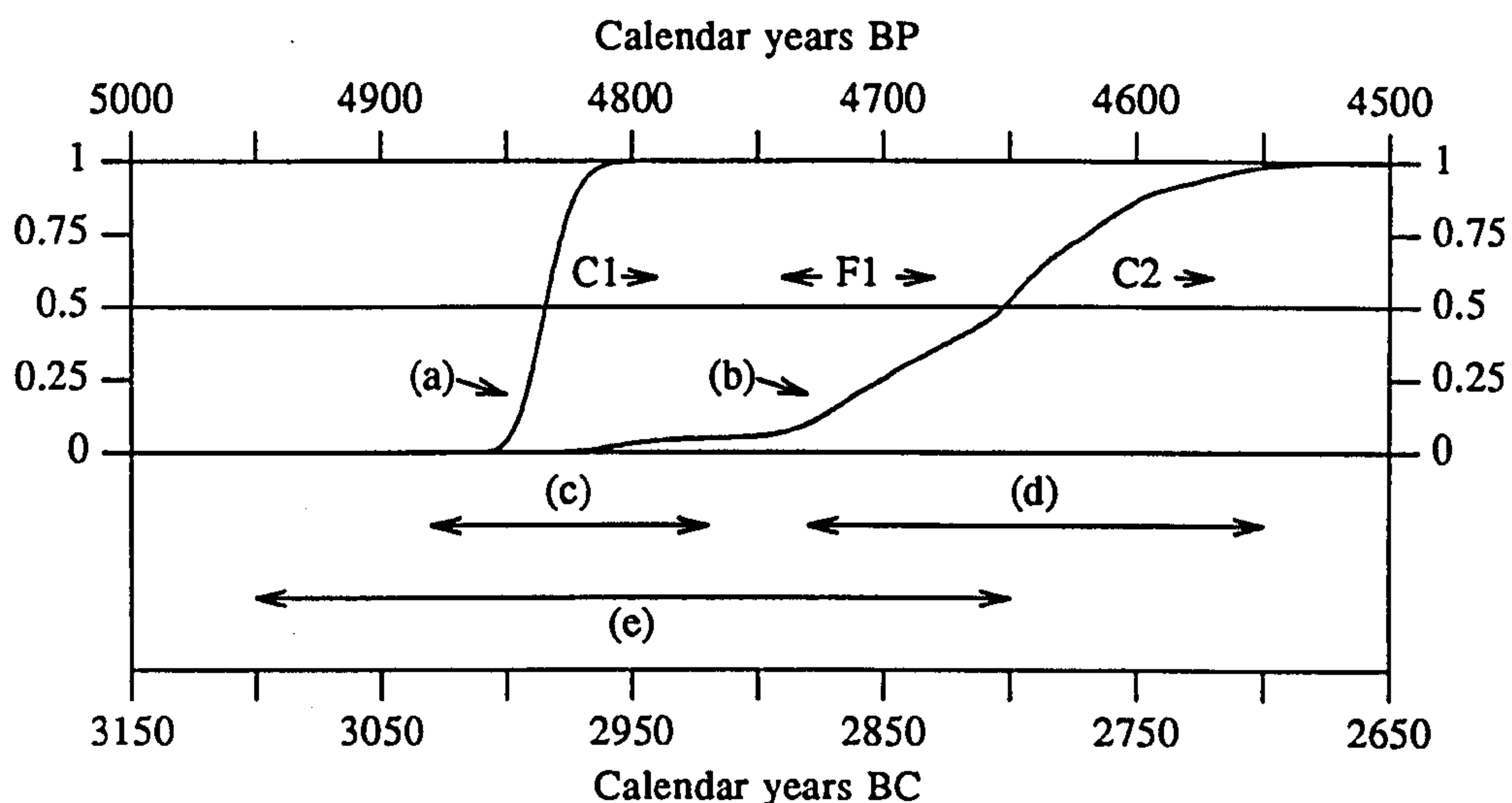


Figure 7.9

Cumulative posterior distribution functions for β (a) and β_3 (b) and the probable position in time (given as ranges) for the events C1, F1, construction and fire of fence (i) and C2, construction of fence (ii). (c) and (d), are estimates of the calendar dates for the fires of fences (i) and (ii) respectively, given by Ottaway (1988). (e) is the estimate of the floruit of the site, as given by Aitchison *et al.* (1991).

probability. Note that two occupation periods could be distinguished within the chronology for the site. This is in relation to the fences (i) and (ii). Broadly speaking, the first occupation period can be estimated as being between 2970 to 2750 BC (between C1 and C2) and the second after 2750 BC (after C2).

Also in Figure 7.9 we have indicated estimates given by other authors. Ottaway (1988) gave as estimates for the dates of fires (i) and (ii) the time ranges 3030 to 2920 BC and 2880 to 2700 BC respectively, see Figure 7.9. Aitchison *et al.* (1991) gave as an estimate for the floruit of the site the time range 3100 to 2800 BC seen in Figure 7.9. In both of these analyses little or no information about the relative chronology of the site was used nor was the relationship between the objects radiocarbon dated and the events of interest clearly established (we have discussed both of these analyses in Section 7.3.1). In particular,

the fact that the determinations cannot be used to date the fires (i) and (ii) (F1 and F2) directly was never considered. In contrast to this, we include such considerations in our statistical analysis stating that the associated calendar years for the determinations predate the construction of the fences and thus predate the fires F1, and F2. The difference in the various approaches can be seen in Figure 7.9 and understood in that perspective. Our estimates are, broadly speaking, 100 to 150 years later than those of Ottaway (1988) or Aitchison *et al.* (1991). Besides this we have not given any precise estimate for the date of the fire (ii). This comes as a result of our robust analysis and of inherent limitations found in the interpretation process.

It now seems clear that the inclusion in the statistical analysis of, what we believe to be, crucial prior information explains the differences between our estimates and those reported by other authors. In this respect, we believe that we have achieved a more realistic interpretation of the Galgenberg dating problem.

7.7 Conclusions

At the beginning of our analysis we stated that the Galgenberg dating problem was to date the fires (i) and (ii). The estimates obtained, though, were not for the dates of the fires themselves, but for the construction of the fences (i) and (ii) (C1 and C2). The fire of fence (i), F1, occurred any time between C1 and C2. According to Figure 7.9, this time range can be broadly estimated as 2950 BC to 2750 BC. To give a more precise range, we would need some information about the time elapsed between the construction and the fire of fence (i). That is, the time elapsed between C1 and F1. For F2 we can only say that it occurred after C2, that is, some time after about 2750 BC. Similarly, in order to give a more precise date for F2 we would need some information about the time elapsed between C2 and F2. For example, it could be argued that the posts in a fence will

deteriorate and may therefore only last, say, 50 years or less. However, we do not have enough knowledge about the Galgenberg site to support this kind of reasoning and therefore the above datings for F1 and F2 represent our best estimates, given present knowledge.

At this point we would like to review what we have achieved so far with the Galgenberg dating problem. Using the published reports for the excavations at the site, we deduced a series of chronological relationships among events observed in the stratigraphy. This gave us a relative chronology for a series of events that, according to our understanding, must have occurred in the site. Given the characteristics of the samples radiocarbon dated we formed four groups from them and deduced the chronological relationships between those groups and the relative chronology of the site.

Using our statistical framework we were able to include in the analysis the relative chronology identified for the site and its 'logical link' with the samples radiocarbon dated. The resulting statistical analysis developed then enabled us to combine the radiocarbon determinations with the relative chronology of the site and thus, in some sense, to date the events observed in the stratigraphy.

Furthermore, a robustness analysis was carried out to investigate how well in accordance the radiocarbon determinations were with the model and prior information used. As a result of this, two determinations were removed as probable outliers and a piece of our prior information reconsidered. We expect that this will give us more robust datings.

Since several archaeological considerations were included in the analysis, the datings obtained could be interpreted in relation to the events observed in the site. Relevant posterior probabilities were calculated and we were able to obtain probable time ranges for these events, on the calendar scale.

As a final comment we would like to mention the factors involved in the analysis.

- (i) A basic model for radiocarbon determinations that allows for the possibility of outliers.
- (ii) Calibration including the variance in the calibration curve.
- (iii) The relative chronology of the site.

Bayes' theorem has allowed us to combine in a consistent manner the above factors with the radiocarbon determinations available and obtain estimates of the dates for the archaeological events of interest. We must remark, however, that such estimates could only be obtained given the assumptions involved in (i), (ii) and (iii). In particular, this is true of the relative chronology assumed for the site which includes the relevant archaeological information available for the dating problem. Indeed, all those assumptions are open to discussion and we do not expect them to be accepted without comment or criticism. Moreover, since one's current knowledge and understanding is always limited, it is likely that further considerations about the archaeology in the site could be used to improve our analysis. Therefore, the dates obtained here can only be believed if our assumptions are accepted.

Chapter 8

Discussion

The topics considered in this thesis may be broadly divided into three groups, namely

- (i) radiocarbon calibration of one radiocarbon determination,
 - (ii) statistical analysis and interpretation of sets of possibly related radiocarbon determinations
- and
- (iii) robustness of our analyses to outliers.

In the light of these, in this Chapter we present some final thoughts for the thesis.

With respect to point (i), we discussed radiocarbon calibration in Chapter 3 where we developed the calibration method used later in the thesis. This represents an original technique for radiocarbon calibration which, as opposed to most other approaches, attempts to take into account the observed variability for the calibration data. We have compared our calibration technique with others and concluded that, when the standard deviations of the determinations to be calibrated are small, it is important to consider that variability.

Our calibration technique takes into account the yearly variation of the atmospheric ^{14}C , $\alpha(\theta)$. Using our method, to calibrate radiocarbon determinations it is necessary to have an estimate for the variance and correlation structure of that process. Since we have access to only small amounts of data relating to the

process $\alpha(\theta)$, we could only obtain broad estimates for this variability. As a result of this, in our examples we used what may well be an overestimate for the variance of $\alpha(\theta)$. Further research is needed to investigate this matter and to establish a more accurate estimate.

In relation to point (ii), in Chapters 4, 5 and 6 we have presented a general statistical framework for the analysis and interpretation of radiocarbon determinations. This framework provides a powerful tool for the statistical analysis of sets of radiocarbon determinations associated with a wide range of dating problems. Using the Bayesian paradigm it is possible to include in the analysis archaeological prior knowledge about the characteristics of the dating problem under study. We have addressed, and proposed solutions to, many common problems encountered in the interpretation of radiocarbon determinations. Furthermore, the framework provides a general approach which is not restricted to the study of specific examples but is applicable to a much wider variety of problems.

We have presented a series of examples that illustrate how our framework is applied. In some cases, despite not using very sophisticated mathematical techniques, we have addressed important dating problems. However, since many problems have large number of unknown parameters, we have used Markov chain Monte Carlo (MCMC) methods to obtain the posterior distributions of interest.

Nevertheless, several questions may arise when applying our framework. A very important issue is how to define reliably the prior distributions involved in the analysis. That is, how should the knowledge elicitation process be carried out? In the examples used in this thesis we have presented various different priors and discussed some alternatives. In particular, in Chapter 7 we have presented a study of how the prior distributions are elicited in a specific dating problem and given some general ideas to tackle the elicitation problem. We acknowledge that, for our approach to work, careful elicitation must be carried out to ensure that the

prior distributions used, correctly reflect the characteristics of the dating problem under study. Therefore, further important research is required to consider the elicitation problems involved in the interpretation of radiocarbon data. This will necessarily involve the users of radiocarbon dating (mainly archaeologists) and thus far-reaching interdisciplinary collaboration.

Another important issue to be considered is the sensitivity of our analyses to changes in prior specification. That is, to what extent our estimates will be affected if the priors used are slightly altered. This is an important problem since, in most real situations, there is not a unique prior that describes the knowledge available about the dating problem under study. It is commonly the case that a set of similar priors could be used. In those circumstances our results should not dramatically change from one prior choice to other. In some of the examples presented in Chapter 5 we briefly discussed changes in prior specification and how the resulting analysis is affected. Further research should not only look at specific examples but try to address the sensitivity problem in a general way.

We used numerical techniques to implement the Bayesian paradigm and obtain the posterior distributions of interest. In most of the cases we have used MCMC methods, more specifically, the Gibbs sampler. This involves large amounts of time both in terms of developing and debugging computer software and CPU. Given the present availability of very powerful computers, however, the need for CPU time is less and less problematic. The complexity of the sampling routines required commonly demands complex software development and a meticulous debugging process. For our approach to be of wide applicability it will be necessary to have well-tested and efficient routines. The routines will involve sampling, integration and data handling. A well-designed environment for developers would be desirable. In any case, a close collaboration between the users of radiocarbon and statisticians will always be required to perform the

analyses and develop the software that may be needed in specific applications.

We need to mention also that there is a disadvantage in using numerical methods. That is, they give little insight into what results should be expected and how the statistical techniques actually work. It would be desirable to obtain a better analytical understanding of the techniques used and to gain insight into how, algorithmically, data and priors are transformed into posterior estimates. However, the chances of attaining such an understanding seem remote given the nature of the likelihood used.

Finally, considering point (iii) above, in Chapters 6 and 7 we have discussed the problem of outliers in radiocarbon data. We believe that we have presented an interesting approach that can successfully be used in a wide variety of problems. Since it is based on a general methodology, our approach is not restricted to specific problems. On the contrary, we have attempted to tackle the problem of outliers in radiocarbon data using an original methodology which has widespread applicability.

Our methodology for identifying outliers in radiocarbon data is based on simple intuitive ideas formalised within a well-defined statistical model. It provides us with interpretable results, obtaining the posterior probability of each determination being an outlier. Here we have presented three examples of how this methodology can be applied. More examples need to be studied of various dating problems and data sets to explore the performance of the methodology.

To model the presence of outliers we have used shifts on the radiocarbon scale. Subsequent research should consider shifts on the calendar scale and, given that, decide upon a clear strategy as to what type of shifts should be used in specific situations. Also, there is the problem of specifying the prior probability of each determination being an outlier. In Chapters 6 and 7 we used a prior

probability of 0.1 of any determination being an outlier. This probability is not definitive and could, perhaps, be reduced if appropriate. On the other hand, in some circumstances, it may be necessary to introduce uncertainty. For example, we could consider a 'beta' prior distribution and thus have a degree of uncertainty for the prior probability of each determination being an outlier.

To conclude, we can say that we have developed an interesting and powerful statistical framework for the Bayesian interpretation of radiocarbon determinations. Using our framework we have been able to tackle, successfully we believe, statistical problems associated with various archaeological problems that in the past have been the source of great confusion. Several further issues need to be addressed, some of which we have outlined above. More examples and theoretical work need to be developed to improve our framework and facilitate its application. Indeed, much else needs to be done.

References

- Abraham, B. and Box, G. E. P. (1978). 'Linear models and spurious observations', *Applied Statistics*, **27**(2), 131-138.
- Aitchison, J. and Dunsmore, I. R. (1975). *Statistical prediction analysis*, Cambridge University Press, Cambridge, UK.
- Aitchison, T. C., Leese, M., Michczynska, D. J., Mook, W. G., Otlet, R. L., Ottaway, B. S., Pazdur, M. F., van der Plicht, J., Reimer, P. J., Robinson, S. W., Scott, E. M., Stuiver, M., and Weninger, B. (1989). 'A comparison of methods used for the calibration of radiocarbon dates', *Radiocarbon*, **31**(3), 846-862.
- Aitchison, T. C., Ottaway, B. S., and Al-Ruzaiza, A. S. (1991). 'Summarizing a group of ^{14}C dates on the historical time scale: with a worked example from the Late Neolithic of Bavaria', *Antiquity*, **65**, 108-116.
- Aitchison, T. C., Ottaway, B. S., and Scott, E. M. (1990). 'Statistical treatment of groups of related radiocarbon dates', in *Second International Symposium ^{14}C and Archaeology*, W. G. Mook and H. T. Waterbolk (ed.), PACT, Groningen, 95-104.
- Aitchison, T. C., Scott, E. M., Harkness, D. D., Baxter, M. S., and , G. T. Cook (1990b). 'Report on stage 3 of the international collaborative program', *Radiocarbon*, **32**(3), 271-278.
- Aitken, M. J. (1990). *Science-based dating in archaeology*, Longman, London.
- Anderson, E. C., Libby, W. F., Weinhouse, S., Reid, A. F., Kirshenbaum, A. D., and Grosse, A. V. (1947). 'Radiocarbon from cosmic radiation', *Science*,

105, 576-577.

Bada, J. L., Gillespie, R., Gowlett, J. A. J., and Hedges, R. E. M. (1984).

'Acceleration mass spectrometry radiocarbon ages of amino acid extracts from California paleoindian skeletons', *Nature*, 312, 442-444.

Baillie, M. G. L. (1990). 'Checking back on an assemblage of published radiocarbon dates', *Radiocarbon*, 32(3), 361-366.

Baillie, M. G. L. and Pilcher, J. R. (1988). 'Make a date with a tree', *New Scientist*, 117, 48-51.

Barfield, L. H. (1991). 'Wesses with and without Mycenae', *Antiquity*, 65, 102-107.

Barnett, V. (1973). *Comparative Statistical Inference*, John Wiley & Sons Ltd., Chichester.

Barnett, V. (1978). 'The study of outliers: Purpose and model', *Applied Statistics*, 27(3), 242-250.

Barnett, V. and Lewis, T. (1978). *Outliers in statistical data*, John Wiley & Sons Ltd., Chichester.

Barnett, V. and Lewis, T. (1984). *Outliers in statistical data (2nd ed.)*, John Wiley & Sons Ltd., Chichester.

Baxter, M. S. (1990). 'Report of the international workshop of intercomparison of radiocarbon laboratories: a summary of the meeting', *Radiocarbon*, 32(3), 389-391.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis (2nd ed.)*, Springer-Verlag Inc., New York.

- Billingsley, P. (1968). *Convergence of probability measures*, John Wiley & Sons, Inc, New York.
- Bowman, S. (1990). *Radiocarbon Dating*, British Museum Publications, London.
- Box, G. E. P. and Jenkins, G. M. (1976). *Time series analysis: forecasting and control*, Holden-day, Oakland, USA.
- Box, G. E. P. and Tiao, G. C. (1968). 'A Bayesian approach to some outlier problems', *Biometrika*, 55(1), 119-129.
- Buck, C. E., Kenworthy, J., Litton, C. D., and Smith, A. F. M. (1991). 'Combining archaeological and radiocarbon information: a Bayesian approach to calibration', *Antiquity*, 65, 808-821.
- Buck, C. E., Litton, C. D., and Smith, A. F. M. (1992). 'Calibration of radiocarbon results pertaining to related archaeological events', *Journal of Archaeological Science*, 19, 497-512.
- Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992). 'Hierarchical Bayesian Analysis of Change Point Problems', *Applied Statistics*, 41(2), 389-405.
- Chauvet, W. (1863). 'Appendix: Method of least-squares', in *A manual on practical and spherical astronomy*, Lippincott (ed.), Philadelphia.
- Clark, R. M. (1979). 'Calibration, cross-validation and Carbon-14. I', *Journal of the Royal Statistical Society A*, 142, 47-62.
- Clark, R. M. and Morgan, R. A. (1983). 'An alternative statistical approach to the calibration of floating tree-ring chronologies: two sequences from the Somerset Levels', *Archaeometry*, 25, 3-16.

- Clark, R. M. and Renfrew, C. (1972). 'A statistical approach to the calibration of floating tree-ring chronologies using Radiocarbon dates', *Archaeometry*, **14**, 5-19.
- Clark, R. M. and Sowray, A. (1973). 'Further statistical methods for the calibration of floating tree-ring chronologies.', *Archaeometry*, **15**, 255-266.
- Clymo, R. S. (1991). 'Peat growth', in *Quaternary landscapes*, L. C. K. Shane, E. J. Cushing (ed.), University of Minnesota Press, Minneapolis, 76-112.
- Clymo, R. S., Oldfield, F., Appleby, P. G., Pearson, G. W., Ratnesar, P., and Richardson, N. (1990). 'The record of atmospheric deposition on a rainwater-dependent peatland', *Phil. Trans. R. Soc. Lond. B*, **327**, 331-338.
- Damon, P. E. (1970). 'Arizona radiocarbon dates for dendrochronologically dated samples', in *Proceedings of the Twelfth Nobel Symposium*, I. U. Olsson (ed.), Wiley and Sons, New York, 615-618.
- Damon, P. E., Cheng, S., and Linick, T. W. (1989). 'Fine and hyperfine structure in the spectrum of secular variations of the atmospheric ^{14}C ', *Radiocarbon*, **31**(3), 704-718.
- Damon, P. E., Donahue, D. J., Gore, B. H., Hatheway, A. L., Jull, A. J. T., Linick, T. W., Sercel, P. J., Toolin, L. J., Bronk, C. R., Hall, E. T., Hedges, R. E. M., Housley, R., Law, I. A., Perry, C., Bonani, G., Trumbore, S., Ambers, J. C., Bowman, S. G. E., Leese, M. N., and Tite, M. S. (1989b). 'Radiocarbon dating of the Shroud of Turin', *Nature*, **337**, 611-615.
- Dawid, A. P. (1973). 'Posterior expectations for large observations', *Biometrika*, **60**, 664-667.

DeGroot, M. H. (1970). *Optimal statistical decisions*, McGraw-Hill, New York.

Devroye, L. (1986). *Non-uniform Random Variate Generation*, Springer, New York.

Fedick, S. L. and Taube, K. A. (1991). 'The rôle of radiocarbon dating in Maya archaeology: four decades of research', *Radiocarbon after four decades: An interdisciplinary perspective*, R. E. Taylor, A. Lang and R. S. Kra (ed.), Springer-Verlag, New York, 403-420.

de Finetti, B. (1974). *Theory of Probability: a critical introductory treatment*, Wiley, New York. 2 vols.

Gelfand, A. E. and Smith, A. F. M. (1990). 'Sampling based approaches to calculating marginal densities', *Journal of the American Statistical Association*, 85, 398-409.

Gelman, A. and Rubin, D. B. (1992). 'Inference from iterative simulation using multiple sequences', *Statistical Science*, 7(4), 457-511.

Geman, S. and Geman, D. (1984). 'Stochastic relaxation, Gibbs distributions and Bayesian restoration of images', *I.E.E.E Transactions of Pattern Analysis and Machine Intelligence*, 6, 721-741.

Geyer, C. J. (1992). 'Practical Markov chain Monte Carlo', *Statistical Science*, 7(4), 473-511.

Gilks, W. R., Clyton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D., and Kirby, A. J. (1993). 'Modelling Complexity: Applications of Gibbs sampling in Medicine', *Journal of the Royal Statistical Society B*, 55(1).

Gillespie, R. (1986). *Radiocarbon User's Handbook*, Oxford University Committee for Archaeology, Oxford.

Gowlett, J. A. J. (1986). 'Problems in dating the early human settlement of the Americas', in *Archaeological results from accelerator dating*, J. A. J. Gowlett and R. E. M. Hedges (ed.), Oxford University Committee for Archaeology Monograph 11, Oxford, 51-59.

Guttman, I., Dutter, R., and Freeman, P. R. (1978). 'Care and handling of univariate outliers in the general linear model to detect spuriousity-A Bayesian approach', *Technometrics*, 20(2), 187-193.

Harrison, R. J. and Wainwright, J. (1991). 'Dating the Bronze Age in Spain: a refined chronology for the high-altitude settlement of El Castillo', *Oxford Journal of Archaeology*, 10(3), 261-268.

Hassan, F. A. and Robinson, S. W. (1987). 'High-precision radiocarbon chronometry of ancient Egypt, and comparisons with Nubia, Palestine and Mesopotamia', *Antiquity*, 61, 119-135.

Hays, W. L. and Winkler, R. L. (1970). *Statistics (probability, inference and decision)*, Holt, Rinehart and Winston, Inc., New York.

Helskog, K. and Schweder, T. (1989). 'Estimating the number of contemporaneous houses from ^{14}C dates', *Antiquity*, 63, 166-172.

Hill, B. M. (1974). 'On coherence, inadmissibility and inference about many parameters in the theory of least squares', in *Studies in Bayesian econometrics and statistics*, Fienberg and Zellner (ed.), North-Holland, Amsterdam, 555-584.

- ISG, (1982). 'An inter-laboratory comparison of radiocarbon measurements in tree rings', *Nature*, **298**, 619-623.
- Jeffreys, H. S. (1961). *Theory of Probability*, University Press, Oxford. Third Edition.
- Jong, A. F. M. De (1981). *Natural ^{14}C variations*, PhD thesis, Rijksuniversiteit te Gröningen, Gröningen.
- Kirch, P. V., Flenley, J. R., and Steadman, D. W. (1991). 'A radiocarbon chronology for human-induced environmental change on Mangaia, southern Cook islands, Polynesia', *Radiocarbon*, **33**(3), 317-328.
- Kirch, P. V., Hunt, T. L., and Tyler, J. (1989). 'A radiocarbon sequence from the Toaga site, Ofu island, American Samoa', *Radiocarbon*, **31**(1), 7-15.
- Law, R. G. (1975). 'Radiocarbon dates from Rangitoto and Motutapu, a consideration of the dating accuracy', *New Zealand Journal of Science*, **18**, 441-451.
- Leach, B. F. (1972). 'Multi-sampling and absolute dating methods: A problem of statistical combination for archaeologists', *New Zealand Archaeological Association Newsletter*, **15**(3), 113-116.
- Leese, M. N. (1988). 'Methods for finding calendar date bands from multiple-valued radiocarbon calibration curves', in *Computer Applications and Quantitative Methods in Archaeology 1986*, C. L. N. Ruggles and S. P. Q. Rahtz (ed.), British Archaeological Reports, International Series, Oxford, **393**, 147-151.
- Lehmann, E. and Scheffé, H. (1950). 'Completeness, similar regions and unbiased estimation', *Sankhya*, **10**, 305-340.

- Libby, W. F. (1946). 'Atmospheric helium three and radiocarbon from cosmic radiation', *Physics Review*, **69**, 671-672.
- Libby, W. F. (1954). 'Chicago radiocarbon dates, IV', *Science*, **119**, 135-140.
- Libby, W. F., Anderson, E. C., and Arnold, J. R. (1949). 'Age determination by radiocarbon content: Worldwide assay of natural radiocarbons', *Science*, **109**, 227-228.
- Lindley, D. V. (1957). 'A statistical paradox', *Biometrika*, **44**, 187-192.
- Litton, C. D. and Leese, M. N. (1991). 'Some statistical problems arising in radiocarbon calibration', in *Computer Applications and Quantitative Methods in Archaeology 1990*, K. Lockyear and S. Rahtz (ed.), Tempus Reparatum, British Archaeological Reports, International Series 565, Oxford, 101-109.
- Long, A. and Kalin, J. (1990). 'A suggested quality assurance protocol for radiocarbon dating', *Radiocarbon*, **32**(3), 329-334.
- Long, A. and Rippeteau, B. (1974). 'Testing contemporaneity and averaging radiocarbon dates', *American Antiquity*, **39**, 205-215.
- Manning, S. W. and Weninger, B. (1992). 'A light in the dark: Archaeological wiggle matching and the absolute chronology of the close of the Aegean Late Bronze Age', *Antiquity*, **66**, 636-663.
- Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the theory of statistics*, McGraw-Hill, Tokyo.
- Naylor, J. C. and Smith, A. F. M. (1988). 'An archaeological inference problem', *Journal of the American Statistical Association*, **83**(403), 588-595.

- Nydal, R. (1989). 'A critical review of radiocarbon dating of a Norse settlement at L'Anse aux Meadows, Newfoundland Canada', *Radiocarbon*, 31(3), 976-985.
- O'Hagan, A. (1979). 'On outlier rejecting phenomena in Bayes inference', *Journal of the Royal Statistical Society - B*, 41(3), 358-367.
- Orton, C. (1980). *Mathematics in archaeology*, William Collins Sons & Co Ltd, London.
- Orton, C. (1983). 'A statistical technique for integrating C-14 dates with other forms of dating evidence', in *Computer Applications and Quantitative Methods in Archaeology*, J. Haigh (ed.), 115-124.
- Otlet, R. L., Walker, A. J., Hewson, A. D., and Burleigh, R. (1980). '¹⁴C interlaboratory comparison in the UK: Experiment design, preparation and preliminary results', *Radiocarbon*, 22(3), 936-946.
- Ottaway, B. S. (1973). 'Dispersion diagrams: a new approach to the display of carbon-14 dates', *Archaeometry*, 15(1), 5-12.
- Ottaway, B. S. (1986). 'Is radiocarbon obsolescent for archaeologists?', *Radiocarbon*, 28(2A), 732-738.
- Ottaway, B. S. (1988). 'The Galgenberg, a late neolithic enclosure in Bavaria', in *Enclosures and Defences in the Neolithic of Western Europe*, C. Burgess, P. Topping, C. Mordant and M. Madison (ed.), British Archaeological Reports, International Series, 403(ii), 391-418.
- Paz, O. (1958). *La estación violenta*, Fondo de Cultura Económica, Mexico D.F..
- Pazdur, M. F. and Krzanowski, A. (1991). 'Fechados radiocarbónicos para los sitios de la cultura Chancay', in *Estudios sobre la cultura Chancay, Perú*,

- Andrzej Krzanowski (ed.), Universidad Jaguelona, Kraków, Poland, 115-132.
- Pazdur, M. F. and Michczynska, D. J. (1989). 'Improvement of the procedure for probabilistic calibration of radiocarbon dates', *Radiocarbon*, 31(3), 824-832.
- Pazdur, M. F., Michczynska, D. J., and Walanus, A. (1990). 'Bayesian approach to probabilistic calibration of radiocarbon ages', in *Second International Symposium ¹⁴C and Archaeology*, W. G. Mook and H. T. Waterbolk (ed.), PACT, Groningen, 69-79.
- Pearson, G. W. (1986). 'Precise calendar dating of known growth-period samples using a 'curve fitting' technique', *Radiocarbon*, 28(2A), 292-299.
- Pearson, G. W., Pilcher, J. R., Baillie, M. G. L., Corbett, D. M., and Qua, F. (1986). 'High-precision ¹⁴C measurements of Irish oaks to show the natural ¹⁴C variations from AD 1840 - 5210 BC', *Radiocarbon*, 28(2B), 911-34.
- Pearson, G. W. and Stuiver, M. (1986). 'High-precision calibration of the radiocarbon time scale, 500 - 2500 BC', *Radiocarbon*, 28(2B), 839-862.
- Pearson, G. W. and Stuiver, M. (1993). 'High-decadal bidecadal calibration of the radiocarbon time scale, AD 1950-500 and 2500-6000 BC', *Radiocarbon*, 35(1), 1-23.
- van der Plicht, J., Mook, W. G., and Hasper, H. (1990). 'Automatic calibration of radiocarbon ages', in *Second International Symposium ¹⁴C and Archaeology*, W. G. Mook and H. T. Waterbolk (ed.), PACT, Groningen, 81-94.
- Polach, H. A. and Golson, J. (1966). *Collections of specimens for radiocarbon dating and interpretation of results*, Australian Institute of Aboriginal Studies (Manual 2), Canberra.

- Raftery, A. and Lewis, S. (1992). 'How many iterations in the Gibbs sampler?', in *Bayesian Statistics 4*, J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith (ed.), Oxford University Press, Oxford, 765-776.
- Ralph, E. K. and Michael, H. N. (1970). 'MASCA radiocarbon dates for sequoia and bristlecone-pine samples', in *Proceedings of the Twelfth Nobel Symposium*, I. U. Olsson (ed.), Wiley and Sons, New York, 615-618.
- Ralph, E. K., Michael, H. N., and Han, M. C. (1973). 'Radiocarbon dates and reality', *MASCA Newsletter*, 9(1), Applied Science Center for Archaeology, University of Pennsylvania.
- Renfrew, C. (1973). *Before Civilization*, Penguin Books Ltd, Harmondsworth, England.
- Renfrew, C. and Clark, R. M. (1974). 'Problems of the radiocarbon calendar and its calibration', *Archaeometry*, 16(1), 5-18.
- Roberts, G. O. (1992). 'Convergence diagnostics of the Gibbs sampler', in *Bayesian Statistics 4*, J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith (ed.), Oxford University Press, Oxford, 777-784.
- Robinson, S. W. (1986). 'A computational procedure for the utilization of high-precision radiocarbon curves', *Open-File Report, United States Department of the Interior Geological Survey*.
- Saville, A., Gowlett, J. A. J., and Hedges, R. E. M. (1987). 'Radiocarbon dates from the chambered tomb at Hazleton (Glos.): a chronology for neolithic collective burial', *Antiquity*, 61, 108-119.
- Scott, E. M., Aitchison, T. C., Harkness, D. D., Cook, G. T., and Baxter, M. S. (1990c). 'An overview of all three stages of the international radiocarbon

intercomparison', *Radiocarbon*, **32**(3), 309-319.

Scott, E. M., Baxter, M. S., Aitchison, T. C., Harkness, D. D., and Cook, G. T. (1989). 'An interim progress report on stages 1 and 2 of the International Collaborative Program', *Radiocarbon*, **31**(3), 414-421.

Scott, E. M., Baxter, M. S., Aitchison, T. C., Harkness, D. D., and Cook, G. T. (1990). 'An overview of some interlaboratory studies', *Radiocarbon*, **32**(3), 259-265.

Scott, E. M., Baxter, M. S., Harkness, D. D., Aitchison, T. C., and , G. T. Cook (1990b). 'Radiocarbon: present and future perspectives on quality assurance', *Antiquity*, **64**, 319-322.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Chapman and Hall, London.

Smith, A. F. M. and Roberts, G. O. (1993). 'Bayesian Computation via the Gibbs Sampler and Related Markov chain Monte Carlo Methods', *Journal of the Royal Statistical Society B*, **55**(1), 3-23.

Smith, R. I. L. and Clymo, R. S. (1984). 'An extraordinary peat-forming community on the Falkland Islands', *Nature*, **309**, 617-620.

Spaulding, A. C. (1958). 'The significance of differences in carbon-14 dates', *American Antiquity*, **23**, 309-311.

Stephens, D. A. (1994). 'Bayesian retrospective multiple change point identification', *Applied Statistics*, **43**(4), 159-178.

Stuiver, M. and Pearson, G. W. (1986). 'High-precision calibration of the radiocarbon time scale, AD 1950 - 500 BC', *Radiocarbon*, **28**(2B), 805-838.

Stuiver, M. and Polach, H. A. (1977). 'Discussion-reporting of ^{14}C data', *Radiocarbon*, 10, 355-363.

Stuiver, M. and Quay, P. D. (1981). 'Atmospheric ^{14}C changes resulting from fossil fuel CO_2 release and cosmic ray flux variability', *Earth and Planetary Sciences Letters*, 1153, 349-362.

Stuiver, M. and Reimer, P. (1986). 'A computer program for radiocarbon age calibration', *Radiocarbon*, 28(2B), 1022-1030.

Stuiver, M. and Reimer, P. J. (1993). 'Extended ^{14}C data base and revised CALIB 3.0 ^{14}C age calibration program', *Radiocarbon*, 35(1), 215-230.

Suess, H. E. (1970). 'Bristlecone-pine calibration of the radiocarbon time scale 5200 B.C. to the present', in *Proceedings of the Twelfth Nobel Symposium*, I. U. Olsson (ed.), Wiley and Sons, New York, 303-311.

Switsur, V. R. (1990). 'A consideration of some basic ideas for quality assurance in radiocarbon dating', *Radiocarbon*, 32(3), 341-346.

de Vries, H. (1958). 'Variation in concentration of radiocarbon with time and location on Earth', *Koninkl. Nederlandse Akad. Wetensch. Proc. ser. B*, 61(2), 1-9.

Ward, G. K. and Wilson, S. R. (1978). 'Procedures for comparing and combining radiocarbon age determinations: a critique', *Archaeometry*, 20(1), 19-31.

Warner, B. G., Clymo, R. S., and Tolonen, K. (1993). 'Implications of peat accumulation at point Escuminac, New Brunswick', *Quaternary Research*, 39, 245-248.

Weninger, B. (1986). 'High-precision calibration of archaeological radiocarbon dates', *Acta Interdisciplinaria Archaeologica*, 4, 11-53.

Willis, E. H., Tauber, H., and Münnich, K. O. (1960). 'Variations in the atmospheric radiocarbon concentrations over the past 1300 years', *American Journal of Science Radiocarbon Supplement*, 2, 1-4.

Wilson, S. R. and Ward, G. K. (1981). 'Evaluation and clustering of radiocarbon age determinations: procedures and paradigms', *Archaeometry*, 23(1), 19-39.