Archaeometry **20**, (1), (1978) 19–31. Printed in Great Britain

PROCEDURES FOR COMPARING AND COMBINING RADIOCARBON AGE DETERMINATIONS: A CRITIQUE

G. K. WARD

Department of Prehistory, Research School of Pacific Studies, The Australian National University, P.O. Box 4, *Canberra*, 2600, *Australia*

and S. R. WILSON

Department of Statistics, Research School of Social Sciences, The Australian National University, P.O. Box 4, *Canberra*, 2600, *Australia*

1. INTRODUCTION

In the search for evidence to evaluate hypotheses concerning a sequence of radiocarbon age determinations, it is usually enlightening to incorporate (relatively) objective procedures of a statistical nature. In doing this, it is necessary to be aware of the underlying model one is using, that is, to be aware of the assumptions one is making, in applying the statistical procedure(s). Then one can judge whether or not the technique is appropriate to the evidence one wishes to acquire.

On perusal of the literature we have noted that the statistical technique used is not necessarily appropriate to the particular situation. In the first instance we have been concerned with techniques for comparing and, in the second, with those for combining a series of radiocarbon age determinations. In this article we consider these two aspects separately for a variety of situations, attempting to show clearly the modelling procedure involved. Methods for simultaneously comparing and combining a series of radiocarbon age determinations (i.e. the clustering of determinations, using appropriate statistical criteria) are considered elsewhere (Wilson and Ward n.d.).

In the following, we discuss some of the statistical techniques that have been proposed in the literature for comparing and combining radiocarbon age determinations; this discussion is set in the context of the model that has been assumed and the appropriateness of this model to the problem. (This is not meant to be an exhaustive review of the literature, but rather an indication of the extent of the confusion that exists.) We will start, however, by making recommendations as to which technique(s) should be applied, according to the model(s) which in our view is/are appropriate, and demonstrate the dependence of the model on several factors, such as the type of material used for dating, its method of collection, and the evidence being sought.

2. CONSIDERATIONS FOR THE COMPARISON AND COMBINATION OF RADIOCARBON DETERMINATIONS

A radiocarbon age determination is usually presented in the form $A \pm E$, where A is the estimate of the radiocarbon age b.p. and E is its standard deviation due to counting error, as supplied by the radiocarbon dating laboratory. Both these estimates are obtained by a

complex procedure which varies from laboratory to laboratory (for a description of the procedure followed at the Australian National University Radiocarbon Dating Laboratory see Polach (1976)).

Ideally one should manipulate age estimates which have been derived by comparable procedures, and which can be shown to have comparable sampling distributions. In the following discussion we shall assume this to be the case. However, while this certainly should be a valid assumption for determinations supplied by the same laboratory, its validity may well be questionable for determinations supplied by different laboratories. Also, we have made the usual assumption that the estimate of the radiocarbon age follows a normal distribution. (Although, as Polach (1976) has written, 'It is not appropriate to pool ages $\pm$ errors. A more accurate procedure is to pool . . . the δC^{14} values, because the "depletion error" is normally distributed round the δC^{14} % value, while the "age error" is lognormally distributed round the Age' b.p. Use of this more accurate procedure makes little difference for age estimates less than approximately 30 000 years b.p., but careful consideration should be given to the assumption of normality for age estimates greater than this value.)

We note that the conventional representation $A \pm E$ actually represents a 68% confidence interval, rather than the more usual statistical representation of $(x-y, x+y)$ representing an approximate 95% confidence interval. In this paper we shall use the notation (A, E^2);* here E^2 is the variance (or squared standard error) of the radiocarbon age estimate, due to imprecision of the measurement. An approximate 95% confidence interval for the radiocarbon age estimate, taking into account only the counting error, is then given by $(A-2E, A+2E)$.

First, suppose one wishes to compare a series of radiocarbon determinations. Then the prime consideration is whether one has two or more determinations made on the same object or different parts of the same object (as occurs for instance when several laboratories are evaluating counting procedures or reference standards or when parts of an object of unknown age are distributed among these laboratories for analysis); or whether one has determinations made upon two or more samples known not to be from the same object or which cannot be assumed to have been derived from the same object. There is a fundamental difference in the sampling considerations and in the error factors in these two situations (referred to now as Case I and Case II respectively).

In the Case I situation, one can assume that all determinations have the same true mean and that differences have occurred due to changes in the circumstances (often uncontrollable) under which the determination was made. The model that we propose as appropriate here, given present-day knowledge, is the following. One observes a series of n radiocarbon determinations $\{A_i, E_i^2; i=1, \ldots, n\}$ which can be taken respectively to be realizations of random variables $\{a_i; i=1, \ldots, n\}$ where each a_i has the same expected value, say θ, i.e.

$$E(a_i)=\theta \qquad i=1, \ldots, n.$$

If one assumes that the only sources of errors are due to the counting procedure and that these errors are comparable, as discussed above (noting that appropriate changes must be made if this is not the case), then one can assume that

* We have noted some confusion concerning the representation of the error factor pertaining to age estimates. Use of the notation (A,E^2) would help avoid this confusion. A major advantage to such a notation is that one then could identify more clearly the components of the variance of the radiocarbon age estimate that have been taken into account, by writing the age in the form $(A,E^2+F^2+G^2)$ (see below for definitions of these additional error factors), whereas it is not immediately clear what is meant by y in the expression $(x \pm y)$.

$$a_i = \theta + e_i \qquad i = 1, \ldots, n,$$

where e_i is assumed to be normally distributed with mean zero and variance E_i^2, i.e.

$$e_i \sim N(0, E_i^2).$$

To test the hypothesis that the series of determinations are consistent (i.e. all have effectively the same age), one determines the pooled mean, A_p, where

$$A_p = \left(\sum_1^n A_i/E_i^2\right) \Big/ \left(\sum_1^n 1/E_i^2\right) \tag{1}$$

and then uses the test statistic, T, given by

$$T = \sum_1^n (A_i - A_p)^2/E_i^2 \tag{2}$$

which has a chi-square distribution on $n-1$ degrees of freedom under the null hypothesis.*

If the determinations are judged not to be significantly different then they can be combined, the pooled age being A_p, given by (1), and the variance of the pooled age being given by

$$\mathrm{V}(A_p) = \left(\sum_1^n 1/E_i^2\right)^{-1}. \tag{3}$$

If the determinations are judged to be significantly different then they should not be combined, but need careful reconsideration. To determine objectively which observation(s) is/are outliers, a clustering type of approach involving the likelihood ratio is recommended and this is discussed elsewhere (Wilson and Ward n.d.).

In recent years there has been an increasing awareness by archaeologists, among other Quaternary researchers, that a single radiocarbon determination does not a secure date make. This is due not only to the 'counting error' but also to unquantifiable errors that may occur during the complex dating procedure with unknown probability. (See Polach (1976) for a discussion of these.) The above analysis suggests that a more secure date may be obtained from two or more determinations (for random samples from the object) rather than from devoting valuable resources to decreasing the counting statistic variance, E^2.

Considering now the Case II situation, one does not *know* whether all determinations are estimating the same date (or effectively indistinguishably different dates).† Also, it is necessary to make the assumption that the group of (carbon) samples from which the determinations are made is a random sample. Suppose one has a (random) sample of determinations of size n $\{(A_1, E_1^2), \ldots (A_i, E_i^2) \ldots (A_n, E_n^2)\}$, then these can be regarded as having true or 'real' ages, $R_1, \ldots, R_i, \ldots R_n$ respectively. One wants, eventually, to make inferences concerning $\{R_i, i = 1, \ldots, n\}$, the real ages corresponding to the observations $\{(A_i, E_i^2), i = 1, \ldots, n\}$. (Compare with Case I, where one knew that the real age was R for *all* samples, and wished to make inferences concerning $\{a_i, i = 1, \ldots, n\}$.) Unfortunately, the distribution of $\{\hat{R}_i\}$, the estimated (or calibrated) real age for R_i, is unknown and is likely to be complex, for it involves not only the errors inherent in the radiocarbon determinations (counting error)

* This is the same test used by Clark (1975, p. 252 and tables 2 and 3). $T = (n-1)F$, where F is given by expression (2) of Law (1975). It should be noted, however, that there F is *not* a variance ratio.

† One does not need *a priori* strong subjective evidence for the null hypothesis to be or not to be true to apply the method given here.

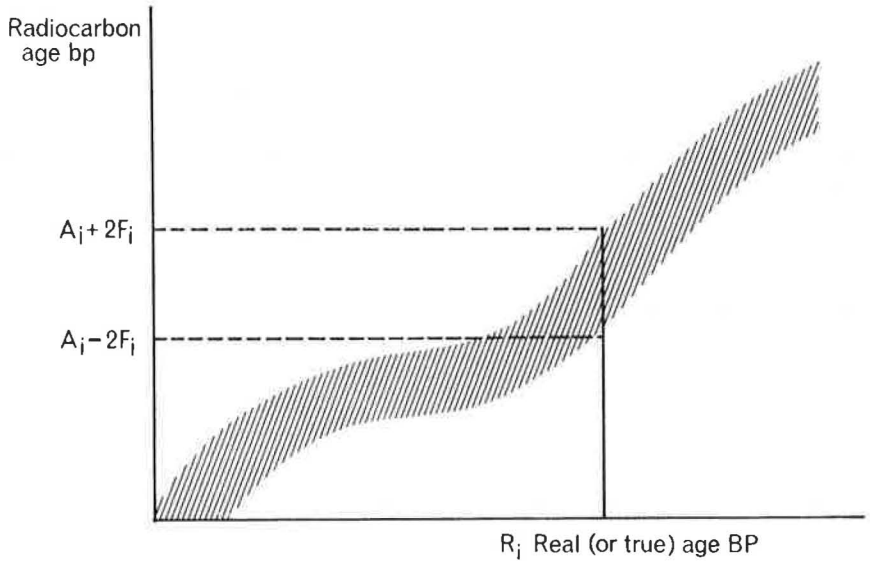

Figure 1 *Schematic relationship between real age* R_i, *of an object and the approximately* 95% *confidence limits for the radiocarbon age, assuming no counting error*

but also, among others (Polach 1976), the unknown error factor in the calibration curve. This error factor is quantifiable, however, and therefore should be taken into consideration when dealing with determinations from different objects. So, in making inferences concerning the true dates, $\{R_i, i=1, \ldots, n\}$, one must consider the distribution of random variables $\{a_i, i=1, \ldots, n\}$ corresponding to the realizations $\{A_i\}$, while *also* taking into consideration the calibration error. To simplify the description of the model we are proposing we shall consider, in the first instance, a single determination A_i, and suppose that $E_i^2=0$. Now the calibration curve is not necessarily unique and may vary for different objects (of different materials and from different locations) of essentially the same age. Then (approximately) 95% confidence limits for the radiocarbon determination, observed to be A_i, for the object of real age, R_i, are given by (A_i-2F_i, A_i+2F_i), where F_i is the standard deviation of the calibration error for the i^{th} observation. This is shown diagrammatically in figure 1, where the shaded band is a representation of the area wherein the calibration curve probably lies. To put this mathematically, we are regarding each radiocarbon determination A_i as having been 'derived' from its corresponding real date R_i, and this is represented by a realization A_i of a random variable a_i, where

$$a_i = \theta_i + f_i, \tag{4}$$

where the expected value of a_i is equal to θ_i, i.e.

$$E(a_i) = \theta_i$$

and where

$$f_i \sim N(0, F_i^2).$$

We are also assuming here that f_i is statistically independent of f_j $(i \neq j)$. The values we shall

use here for F_i^2 are those given by Clark (1975, p. 256; his s_2^2) of 50^2 and 60^2 years depending on whether the determination is less than or greater than 2700 years b.p. These values are based upon analysis of wood charcoal and dendrochronological relationships, and the possibility that they could be underestimates of the values for material of differing types is under consideration.

Then when one includes the uncertainty in the radiocarbon determination due to the counting error the model (4) becomes

$$a_i = \theta_i + e_i + f_i^* \qquad i = 1, \ldots, n \tag{5}$$

where e_i involves the same assumptions detailed above for Case I and it is also assumed here that e_i and f_i are independent of one another for all i. If there is additional inaccuracy due to the 'sunspot effect', then the model given by (5) becomes

$$a_i = \theta_i + e_i + f_i + g_i^* \qquad i = 1, \ldots, n$$

where g_i is assumed to be normally distributed

$$g_i \sim N(0, G_i^2)$$

and represents the inaccuracy due to this effect. The value of G_i^2 is 70^2 years b.p. where the effect is suspected (and zero otherwise). A further assumption is made here that the set of variables $(e_i, f_i, g_i;\ i = 1, \ldots, n)$ are independent of one another for each observation and from one observation to another. (This is essentially the extension to n observations of the model proposed by Clark (1975) for a single observation, and the correspondence in notation is $s_1^2 \equiv E^2$, $s_2^2 \equiv F^2$, $s_3^2 \equiv G^2$.)

To test the hypothesis that the estimates of the real dates are equal, one first determines the pooled mean, A_p, where

$$A_p = \left(\sum_1^n A_i/S_i^2\right) \Big/ \left(\sum_1^n 1/S_i^2\right) \tag{6}$$

where $S_i^2 = E_i^2 + F_i^2 + G_i^2$ (this corresponds to (1) for Case I). Then one calculates the test statistic T' (corresponding to T given by (2) for Case I) given by

$$T' = \sum_1^n (A_i - A_p)^2/S_i^2 \tag{7}$$

which has a chi-square distribution with $n-1$ degrees of freedom under the null hypothesis

$$H_0: R_1 = \ldots = R_i = \ldots = R_n.$$

If the estimates of the real dates are judged not to be significantly different *and*, if from archaeological considerations, it is deemed appropriate, then the radiocarbon determinations can be combined, the pooled radiocarbon age being A_p given by (6), and the variance of the pooled age being given by

$$\mathrm{V}(A_p) = \left(\sum_1^n 1/S_i^2\right)^{-1}. \tag{8}$$

[It should be stressed that $\mathrm{V}(A_p)$ is the variance of the *mean* of the group of n observations and *not* the variance of the group of determinations. The determination of the variance of

* This type of model is known as a random effects model in the statistical literature.

the group of determinations is of a more complex nature and is considered elsewhere (Wilson and Ward n.d.). The estimated real age and the confidence interval for this real age can be determined by reading off the appropriate values from the calibration curve as shown by Clark (1975) ($V(A_p)$ here is equal to s^2 in his notation).]

If the assumption of independence of, say, f_i and f_j $(i \neq j)$ is relaxed (and this might be more appropriate if one is considering, say, material of the same type and one is using the same calibration curve, or if one is considering Case I) then the test statistic T or T' is of the form

$$T'' = A_c^T V^{-1} A_c$$

where the vector

$$A_c = \begin{bmatrix} A_1 - A_p \\ \cdot \\ \cdot \\ \cdot \\ A_n - A_p \end{bmatrix}$$

where A_p is given by either (1) or (6) and V is the estimate of the variance–covariance matrix of $\{a_i, \ldots, a_n\}$. (T'' simplifies to T and to T' for Case I and Case II respectively, if one assumes f_i is independent of f_j.) A model of this form has been proposed by Clark (n.d.) giving a test statistic of the form T''. His model basically assumes that there is a true calibration curve, for which a close approximation is available from recent research. One then uses the form of this best approximation to determine V, and the off-diagonal terms will vary depending on the approximation used.

3. CRITIQUE OF PREVIOUS APPROACHES

In the following we present a critique of the relevant literature that involves the major attempts at solutions to have been proposed for the problems of comparison and combination of radiocarbon determinations. For each of the selected articles we have attempted to determine the authors' objectives, and to assess their solutions with respect to the objective, the type of Case (I and/or II) and the other considerations given in the previous section. Where deemed necessary, we propose solutions that we consider to be more appropriate and, in so doing, hope to be able to show how the difficulties associated with interpretation of and with inference from a set of radiocarbon determinations are overcome by a clear formulation and analysis of the appropriate model. In addition, we have taken data from this literature to use in simple paradigms for the methods proposed in the previous section.

The first article to consider the type of situation with which we have been concerned appears to be that by Libby (1954) (although it could be adjudged not entirely suitable for comment, since dating procedures have much improved in accuracy and standard errors should no longer be obtained by just taking 'the square root of the total number of counts taken' (1954, p. 136)). It is interesting to note that he used the procedure we recommend above when discussing Case I, for the determination of a more secure date. Namely, he split the sample into (three) parts and made determinations on each part separately. Libby also determined weighted averages as we recommend above for Case I, but he recommended that 'It is probably better, however, to take the arithmetical average . . .' (1954, p. 136). However,

this calculation would assume that the variances of the counting statistics are equal. That this is not so can be shown by calculating

$$F = E^2_{\max}/E^2_{\min} \tag{9}$$

where $E_{\max}$ is the largest value of $E_1, \ldots, E_n$ and $E_{\min}$ is the smallest, and F is distributed according to an F-distribution with (ν_1, ν_2) degrees of freedom, where ν_1 is the number of 'observations' from which $E_{\max}$ is obtained and, similarly, ν_2 from $E_{\min}$. If the variances are equal (and taking into account that the number of 'observations' for each determination here would be effectively infinite) then $F=1.0$. The value one obtains for Libby's data given (p. 136) as $(4.029, 0.05^2)$, $(4.085, 0.07^2)$ and $(4.156, 0.13^2)$ is

$$F = 0.13^2/0.05^2 = 6.76.$$

So the conditions for the arithmetical procedure to be valid are not satisfied and this procedure is not to be recommended.

Another early article is that by Spaulding (1958), wherein he considered both Case I and Case II types of data, without differentiation, and he analysed both cases as if they were of Case I type. Also, he used an analysis of variance technique to compare the dates, assuming that 'the variance of the individual dates [can be] treated as means of samples with an infinite number of observations in each sample' (1958, p. 310). One basic assumption of the ANOVA technique, however, is that the variances within a group, corresponding here to the variance of the individual dates, are equal. If one makes Spaulding's assumption that the variances of the individual determinations are based on an infinite number of observations, then application of the F-test given by (9), (as demonstrated above) shows that the basic assumption of the ANOVA technique does not hold for Spaulding's data. If this assumption was satisfied then in this type of situation the ANOVA result would be *identical* to the result given by T in (2) for Case I and T' in (7) for Case II.

Again, Polach and Golson (1966) do not differentiate between Case II and Case I (their article, however, is concerned basically with Case I data). To determine whether a series of more than two determinations for a single object (i.e. Case I) is consistent they consider a series of pairs of differences. The interpretation of the results for a series of pairs of differences is not straightforward, since the distribution of the results for such a series is awkward due to the lack of independence between some of the pairs (see comments on Polach's 1972 paper below). Again, application of T in (2) is appropriate and, considering the data published therein (p. 16) on three independent measurements on a single piece of wood $(4330, 190^2)$, $(4560, 210^2)$, $(4940, 300^2)$, it is found that

$$A_p = 4525$$

$$\mathrm{V}(A_p) = 128^2$$

and

$$T = 2.99 < 5.99 = \chi^2_{2;\,0.05},$$

and there is no evidence to reject the null hypothesis that the observations are consistent.

In another article Polach (1969) is concerned with Case I situations and gives the formulae corresponding to (1) and (3) for the special case of a sample of size two.

Leach (1972, p. 113) noted that 'some confusion as to the statistical meaning of absolute age statements' exists, and pointed out that the standard error supplied by the laboratory is

basically the standard error of the mean of the 'sample' and *not* the standard error of a population from which the 'sample' has been drawn (but in this situation, the formulae σ_x and $\sigma_{\bar{x}}$ are misleading for the discussion and should be ignored). Leach does not differentiate between Cases I and II and treats both as if they are Case I. He also gives a 'best estimate of age' to be

$$A = (\Sigma A_i/E_i)/(\Sigma(1/E_i)) \quad \text{(our notation)}.$$

However, the model that would provide such an estimate is not stated and hence the values given by (1) for Case I models and (6) for Case II models are recommended here. He states also that a 'best estimate of standard error' is obtained from

$$E = \{\Sigma(A_i - A)^2/n(n-1)\}^{\frac{1}{2}}.$$

This expression, however, is applicable *only* to a special type of model. Here E is the appropriate form of the estimate of the standard deviation of the *arithmetic mean* of a set of single observations from a single population with unknown variance, say σ^2. In the case being considered, however, there is a set of mean values $\{A_1, \ldots, A_n\}$ each with respective standard deviations $\{E_1, \ldots, E_n\}$. The appropriate form for the estimate of the variance of the weighted or pooled mean is given by (3) for Case I models and by (8) for Case II models. [Perhaps it is appropriate to note here that we show elsewhere (Wilson and Ward n.d.) that if the null hypothesis is that each real age R_i comes from the same distribution with a certain mean, in such a way that we have a realization of a random variable a_i where $a_i \sim N(\theta, S_i^2 + \sigma^2)$ $i = 1, \ldots, n$ where θ and σ^2 are unknown, and *if also* $S_1^2 = \ldots = S_n^2 = S^2$ then the estimate of θ is

$$A_p = \sum_1^n A_i/n$$

$$\mathrm{V}(A_p) = \Sigma(A_i - A_p)^2/n(n-1)$$

and the estimate of σ^2 is $\hat{\sigma}^2$ where

$$\hat{\sigma}^2 = \sum_1^n \{(A_i - A_p)^2/n - 1\} - S^2.]$$

It should be noted here that Law (1975) bases his results (for Case II data) on some of Leach's proposals, and he also comments (p. 448) 'No error has been introduced to allow for the uncertainty of the correction curve, as it serves no purpose where the corrected dates are only for comparison with each other'. We hope that we have demonstrated in section 2 the fallacy of this approach.

In considering Case I situations, Polach (1972) first considers 'the significance of the difference between two C^{14} age determinations' (Appendix 1) by determining

$$z = (A_1 - A_2)/(E_1^2 + E_2^2)^{\frac{1}{2}} \tag{10}$$

where $z \sim N(0,1)$. (This test can be derived also under the broader assumption that the two samples are from different populations with different variances.) Polach then notes (1972, p. 703) that 'the difference between two determinations of the same sample is often discussed in terms of their common standard deviation as $(A_1 - A_2)$/S.D.(m) where S.D.(m) is $(E_1^2 + E_2^2)^{\frac{1}{2}}/2$. Hence S.D.($m$) $= z \times \sqrt{2}$'. (Replacing his symbols by ours for clarity here.) He then assesses 'the frequency distribution of pairwise levels of agreement' since he claims

(1972, p. 696), without justification, that 'the expected parent frequency distribution is normal with standard deviation S.D.$(m)=z\times\sqrt{2}$'. This procedure is wrong due to either typographical or algebraic error. To show the latter, first

$$\text{S.D.}(m) = (E_1^2+E_2^2)^{\frac{1}{2}}/2 \text{ by definition}$$
$$= 2z/(A_1-A_2) \quad \text{from (10)}.$$

Secondly, S.D.(m) is the standard deviation of the unweighted, or arithmetical, mean $(A_1+A_2)/2$ rather than that of the weighted mean. More importantly, consider, by way of example, just two pairs of comparisons A_1-A_2 and A_2-A_3; if A_1, A_2 and A_3 are independently and normally distributed then, under the null hypothesis that the determinations are consistent, the vector of comparisons

$$v = \begin{bmatrix} A_1-A_2 \\ A_2-A_3 \end{bmatrix}$$

has a bivariate normal distribution with mean vector $\binom{0}{0}$ and with variance–covariance matrix

$$M = \begin{bmatrix} E_1^2+E_2^2 & -E_2^2 \\ -E_2^2 & E_2^2+E_3^2 \end{bmatrix}.$$

To test the null hypothesis, the appropriate form of the test statistic is

$$S'' = v^T M^{-1} v$$

and under the null hypothesis S'' has a chi-square distribution with two degrees of freedom. This approach obviously can be extended to more than two comparisons. An equivalent, but more straightforward, approach to using S'' is the repeated application of T given by (2) for Case I models. To demonstrate this simply we have extracted some dates from Polach (1972) and given these in table 1. These include three determinations of a single age by the Australian National University laboratory (ANU-7) and another determination of the same age made elsewhere (W-1571). We have included five determinations made on

Table 1 *Data extracted from Polach* (1972, table 3)

Sample number	A_i *Determinations as reported*	E_i^2
ANU-7	14 550	270^2
ANU-7	15 000	600^2
ANU-7	13 700	300^2
W-1571	14 650	500^2
ANU-5	11 700	260^2
C-800	10 860	410^2
L-698D	11 840	100^2
FSU-3	11 245	450^2
Tx-44	10 700	210^2

another group of samples in the second part of the table. To check consistency of the three ANU-7 determinations we obtain, using (1), (3) and (2),

$$A_p = 14\,253$$
$$\mathrm{V}(A_p) = 190^2$$
$$T = 6.16 \quad \text{(compared with } \chi^2_{2;\,0.05} = 5.99),$$

a borderline case concerning consistency of the determinations. However, if we include W-1571, noting the assumption discussed in section 1 that the distributional properties for different procedures are comparable, we obtain

$$A_p = 14\,303$$
$$\mathrm{V}(A_p) = 178^2$$

and

$$T = 6.71 \quad \text{(compared with } \chi^2_{3;\,0.05} = 7.81), \tag{11}$$

from which one would conclude that there is no evidence that the four determinations are not consistent.

Now considering the five determinations of another age, we obtain

$$A_p = 11\,593$$
$$\mathrm{V}(A_p) = 82^2$$

and

$$T = 28.15 \quad \text{(compared with } \chi^2_{4;\,0.05} = 9.49). \tag{12}$$

To test overall consistency of the determinations of both groups of data (or all groups if considering more than two), the chi-square test statistic values (here two values given by (11) and (12) can be combined (since they are independent) to give an overall chi-square value (here 34.86) which can be compared with a chi-square distribution (with, in this instance, seven degrees of freedom ($\chi^2_{7;0.05} = 14.07$)). To determine which observation(s) is/are likely to be outlier(s), the interested reader is referred to the paper by Wilson and Ward (n.d.).

An excellent discussion of the conditions under which averaging is appropriate is given by Long and Rippeteau (1974) who warn against uncritical averaging of determinations. They then question whether the calibrated determinations should be averaged, or whether the uncalibrated values should be averaged and the average calibrated. As we discussed in section 1, considering the usual distributional assumptions and present-day knowledge, the uncalibrated determinations should be averaged, taking into account the eventual calibration for Case II. Long and Rippeteau also propose use of Chauvenet's criterion for the rejection of outliers, but the inappropriateness of this criterion for this situation has been discussed elsewhere (Renfrew and Clark 1974). To determine whether outliers are present, T ((2) Case I) or T' ((7) Case II) should be used. To show the use of T or T' for determining the presence of outliers and the inappropriateness of Chauvenet's criterion, consider the determinations in table 2 used to estimate the age range of the Lamoka Lake site by Long and Rippeteau (1974, p. 22, table 5). The analysis provides a paradigm of the methods proposed here. This set of data is also graphed in figure 2, where the double line

Table 2 *Lamoka Lake site determinations from Long and Rippeteau* (1974, p. 212)

Sample number	*Determinations (bc)**	A_i** *(bp)*	E_i^2*	F_i^2***	S_i^2
C-288	2419	4369	200^2		
M-26	2485	4435	200^2		
Combined					
C-288, M-26****		4402	141^2	60^2	153^2
C-367	3433	5383	250^2	60^2	257^2
M-195	2575	4525	200^2	60^2	209^2
M-911	2521	4471	150^2	60^2	162^2
M-912	2451	4401	125^2	60^2	139^2
Y-1279	2550	4500	80^2	60^2	100^2
Y-1280	2540	4490	80^2	60^2	100^2

* Determinations and error factors as reported by Long and Rippeteau
** In conversion from bc to bp values it was assumed that p = 1950 in all cases
*** Based upon values recommended by Clark 1975
**** See text for explanation

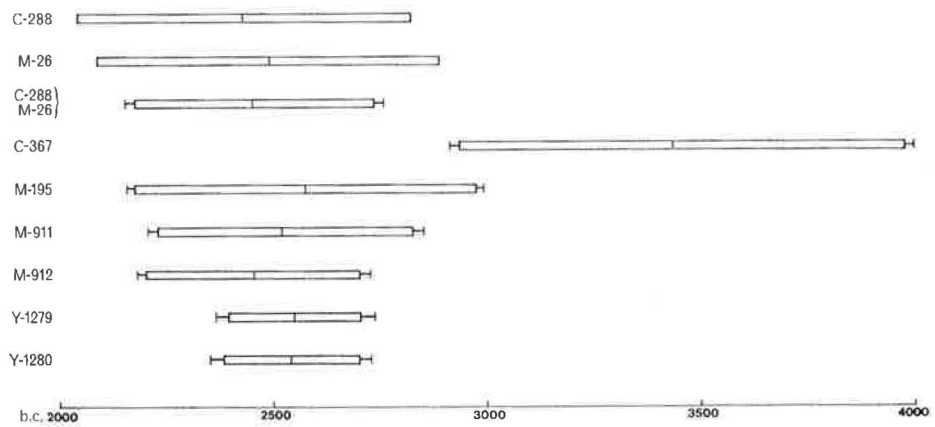

Figure 2 *95% confidence intervals for radiocarbon determinations of table* 2

gives the 95% confidence interval involving just the counting error factor and the extended line gives the 95% confidence interval when the calibration error is taken into account, where appropriate.

Long and Rippeteau reject C-288 using Chauvenet's technique. However, C-288 and M-26 are in fact 'the same sample run by different laboratories' (1974, p. 212) so the procedure for a Case I situation is applicable. From (1) and (2) we obtain

$$A_p = 4402$$

and

$$T = 0.055 < 3.84 = \chi^2_{1;\,0.05}$$

so we have *no* statistical evidence to doubt the consistency of the two determinations, and they may be combined to give the mean value for that object of (4402, 141^2). This value can then be used in the Case II situation to determine whether the estimates of the real dates for the series are the same or not. For the seven age determinations (assuming no sunspot effect) we obtain

$$A'_p = 4507$$
$$V(A'_p) = 52^2$$
$$T' = 12.76 \quad \text{(compared with } \chi^2_{6;\,0.05} = 12.59\text{)}.$$

If, upon subjective or objective evaluation, C-367 is judged to be 'clearly aberrant' (1974, p. 212) then removing C-367 we obtain

$$A'_p = 4472$$
$$V(A'_p) = 53^2$$

and

$$T = 0.65 \quad \text{(compared with } \chi^2_{5;\,0.05} = 11.07\text{)}.$$

An objective method of statistically determining aberrant values is given elsewhere (Wilson and Ward n.d.).

To test 'non-coevalness of radiocarbon dates', Long and Rippeteau (1974, p. 210) first recommended the correction (or calibration) of dates (this has been discussed above and is not recommended) followed by application of the *F*-test as proposed by Spaulding (1958) (also commented on above). They use also an estimate of the variance of the form proposed by Leach (1972) and this too has been commented on previously. Again, the appropriate procedure has been given in the previous section, where the model assumed has been clearly formulated and analysed.

CONCLUSION

That the importance of explicit modelling showing the statistical logic and techniques applied in any attempt to compare and combine age estimates cannot be over-emphasized is argued in this presentation of recommended procedures.

In the first section of this paper the modelling of the recommended procedures was made explicit. Two situations requiring different statistical modelling due to fundamental differences in the types of data were recognized: first, where two or more determinations are obtained from samples of the same object (Case I) and, secondly, where two or more determinations are made from samples which cannot be assumed to derive from the same object (Case II). The latter case applies in most archaeological situations. Inherent in the modelling of either case is the attempt to test the hypothesis that each of a series of determinations provides essentially the same value. The (chi-square) test statistic, T or T', provides a test of this hypothesis. Where the members of a series are found statistically to be insignificantly different from one another, and where archaeological criteria allow, a pooled mean, A_p may be calculated with a new variance, $V(A_p)$, for the *mean* of the grouped determinations. The explicit modelling for deriving these formulae and the procedures for

their application clarify the assumptions made and the techniques used in arriving at objective decisions regarding comparison and combination.

Such modelling aspects usually have been neglected in previous attempts at solution of the problems of comparing and combining a series of radiocarbon determinations. The second part of this paper provides, first, a critique of some previously applied methods again emphasizing the modelling and their appropriateness implicit in their use and, secondly, provides paradigms to exemplify in their appropriate application to archaeological situations the procedures recommended here.

ACKNOWLEDGEMENTS

We are grateful to R. M. Clark of Monash University for his stimulating discussions in this area and to H. A. Polach and M. J. Head of the A.N.U. radiocarbon laboratory.

REFERENCES

Barnard, N., (ed.), 1976, *Scientific methods of research in the study of ancient Chinese and Southeast Asian metal artefacts: a symposium*, Melbourne, National Gallery of Victoria.

Clark, R. M., 1975, A calibration curve for radiocarbon dates, *Antiquity* **XLIX,** 251–266.

Clark, R. M., n.d. Kinks calibration and carbon-14: statistical problems in radiocarbon dating, submitted to Jour. Royal Statistical Soc., Series A.

Law, R. G., 1975, Radiocarbon dates for 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.

Libby, W. F., 1954, Chicago radiocarbon dates IV, *Science* **119,** 135–140.

Long, A. and Rippeteau, B., 1974, Testing contemporaneity and averaging radiocarbon dates, *American Antiquity* **39** (2), 205–215.

Polach, H. A., 1969, Optimisation of liquid scintillation of radiocarbon age determinations and reporting of ages, *Atomic Energy in Australia* **12** (3), 21–28.

Polach, H. A., 1972, Cross checking of NBS oxalic acid and secondary laboratory radiocarbon dating standard. In Rafter and Grant-Taylor (compilers), 1972, **11,** 688–717.

Polach, H. A., 1976, Radiocarbon dating as a research tool in archaeology: hopes and limitations. In Barnard (ed.) 1976, 255–298.

Polach, H. A. and Golson, J., 1966, *Collections of specimens for radiocarbon dating and interpretation of results*. Canberra, Australian Institute of Aboriginal Studies (Manual 2).

Rafter, T. A. and Grant-Taylor, T. (compilers), 1972, *Proceedings of the Eighth International Conference on Radiocarbon Dating*, Wellington: Royal Society of New Zealand (Bulletin 10).

Renfrew, C. and Clark, R. M., 1974, Problems of the radiocarbon calendar and its calibration, *Archaeometry* **16** (1), 5–18.

Spaulding, A. C., 1958, The significance of differences in carbon-14 dates, *American Antiquity* **23,** 309–311.

Wilson, S. R. and Ward, G. K., n.d., Evaluation and clustering of radiocarbon age determinations: Procedures and paradigms (Submitted for publication).