Dealing with Outliers and Offsets in Radiocarbon Dating

Christopher Bronk Ramsey
Research Laboratory for Archaeology, University of Oxford,
Dyson Perrins Building, South Parks Road, Oxford, OX1 3QY

February 9, 2009

Abstract

The wide availability of precise radiocarbon dates has allowed researchers in a number of disciplines to address chronological questions at a resolution which was not possible 10 or 20 years ago. The use of Bayesian statistics for the analysis of groups of dates is becoming a common way to integrate all of the radiocarbon evidence together. However, the models most often used make a number of assumptions which may not always be appropriate. In particular, there is an assumption that all of the radiocarbon measurements are correct in their context and that the original radiocarbon concentration of the sample is properly represented by the calibration curve.

In practice in any analysis of dates some are usually rejected as obvious outliers. However there are Bayesian statistical methods which can be used to perform this rejection in a more objective way (Christen, 1994b) but these are not often used. This paper discusses the underlying statistics and application of these methods, and extensions of them, as they are implemented in OxCal v4.1. New methods are presented for the treatment of outliers, where the problems lie principally with the context rather than the radiocarbon measurement. There is also a full treatment of outlier analysis for samples which are all of the same age which takes account of the uncertainty in the calibration curve. All of these Bayesian approaches can be used either for outlier detection and rejection or in a model averaging approach where dates most likely to be outliers are downweighted.

Another important subject is the consistent treatment of correlated uncertainties between a set of measurements and the calibration curve. This has already been discussed by Jones and Nicholls (2001) in the case of marine reservoir offsets. In this paper the use of a similar approach for other kinds of correlated offset (such as overall measurement bias or regional offsets in the calibration curve) is discussed and the implementation of these methods in OxCal v4.0 is presented.

1 Introduction

Before considering statistical methods for the treatment of outliers and offsets, it is important to understand the underlying mechanisms and issues. There are essentially four main reasons why in any context radiocarbon dates might not give the 'right' result:

- The radiocarbon measurement of a particular sample might not be correct (s)
- The radiocarbon ratio of a sample might be different from that of the associated reservoir (r)
- A whole set of radiocarbon measurements might be biased in some way relative to the calibration curve - either because the measurements themselves are biased or because the reservoir from which the sample draws its carbon might not have the expected radiocarbon isotope ratio (d)
- The sample measured might not relate to the timing of the event being dated (t)

Ideally the uncertainty quoted in the radiocarbon measurement covers the first possibility (s) - though in some instances it may be that the errors in the measurements are not normally distributed. For this reason it might be that in radiocarbon calibration, rather than adopting a Normal distribution, for a more robust model we should be using a longer-tailed distribution (such as a student's t-distribution). Another approach has been suggested where we assume that in a small proportion of cases the measurement effectively has an uncertainty which is larger by some factor (Christen, 1994b, 2003).

In the second case (r), the measurement is correct but the radiocarbon isotope ratio might be different from that of the calibration curve at the associated age for some reason. This might be due to short-term fluctuations in radiocarbon concentrations in particular reservoirs or due to an admixture of carbon from different sources. Such offsets are analogous to the first category but will not be related in any way to the measurement uncertainty or be improved by multiple measurement.

The two categories given for the third reason (d) are very different in their cause, but essentially the same in their effect. The situation here is that the measurements made for the calibration curve and those for the sample have a systematic offset relative to one another. Where such offsets are recognised they can be taken into account using a ΔR correction (Stuiver and Braziunas, 1993). In principle unknown offsets can be treated in similar way using a ΔR with a mean of zero and an uncertainty which reflects the possible scale of offsets between the measurement sets. The correct statistical treatment of such systematic offsets has been described by Jones and Nicholls (2001).

Finally we come to, what is probably the most common form of outlier, where the sample does not for some reason relate to the dated event in the intended way (t). Here the radiocarbon measurements are correct and the values relate correctly to the calibration curve data-sets but there is some sort of calendar offset between the measurement and the event of interest. In some instances such outliers are due to aspects of the deposition process which are hard to understand. In other cases we know why samples might be (or are) outliers: for example in the case of charcoal we usually expect the samples to be older than their context. This type of outlier is not restricted specifically to radiocarbon dating or contamination at the sampling stage.

All of these types of outlier can be treated statistically using essentially similar methods but in slightly different ways. The purpose of this paper is to summarise these approaches and describe the implementation of their algorithms in v4.1 of the analysis program OxCal (Bronk Ramsey, 1995, 2001, 2008).

2 Treatment of outliers

In general there are two main ways of dealing with outliers. The first is to try to identify all outliers and then eliminate them manually from the analysis. If this is possible, then it is probably the best approach since it is then entirely clear what data are being used to support the analysis. The other approach is to assume that we can never really be sure whether any particular measurement is an outlier, but to weight samples according to how likely they are to be correct in a model averaging approach. This outlier analysis approach requires us to provide a prior probability for how likely any individual measurement is to to be wrong and then some model to determine how we should revise this in the context of all of the other information available.

2.1 Manual rejection

How you identify outliers for rejection is a complex topic. The most important considerations are the sample context and details of the measurement process itself. These should allow us to identify which samples might have give anomalous radiocarbon measurements or have a complex depositional history. It is also possible to use statistical methods to indicate which samples seem anomalous within their context to support these decisions. You can either use the outlier analysis methods outlined by Christen (1994b) and in this paper or use the agreement index (Bronk Ramsey, 1995) calculated by OxCal. In practice for identification purposes both methods work well - indeed in almost all instances the same samples will be identified by either. In both cases the level at which we start to reject samples is somewhat arbitrary. If you use the agreement index method, unless a sample has been rejected, all measurements are given equal weight. With outlier analysis samples are progressively down-weighted as they are more likely to be outliers and so the results from the analysis are essentially an average between a model in which the measurement is accepted and one in which it is rejected. If you do not wish to have model averaging, but do wish to use outlier analysis solely for outlier detection, you should first run a model with outlier analysis, see which measurements are likely to be spurious and then run it again, without outlier analysis but with some of the spurious results removed entirely.

In OxCal v4.1 there are three tools which can be used to help with the manual elimination of outliers. The first is the calculation of the agreement index for each sample - if this falls below 60%, rejection should be considered. However, it should be remembered that approximately one in twenty samples are likely to fall below this level and such rejection should also be based on other criteria. Secondly an overall agreement index is calculated A_{model} and if this is above 60% it probably indicates that there is no problem with the model as a whole (and therefore no samples need be rejected). Finally there is a command `Outlier()` which can be used to flag a measurement as a definite outlier and remove it from the model (note in earlier versions of OxCal this command

was `Question`() but otherwise worked in the same way).

2.2 Outlier analysis

In order to deal with outliers statistically we need to have some sort of a model for how we expect them to be distributed. We also ought to define, in the case of radiocarbon, whether we think it is the radiocarbon measurement that is incorrect for some reason or if it is the context that is uncertain. Usually in radiocarbon dating we assume that a specific radiocarbon measurement r_i, will differ from the prevailing radiocarbon concentration $r(t_i)$, given to us by the calibration curve, by an amount ϵ_i such that:

$$r_i = r(t_i) + \epsilon_i \quad (1)$$
$$\epsilon_i \sim N(0, s_i^2 + (s(t_i))^2) \quad (2)$$

So that the difference is entirely accounted for by the uncertainties quoted for the measurement (s_i) and the calibration curve ($s(t_i)$). In outlier analysis we need to be able to deal with other kinds of offset. We introduce another parameter ϕ_i which is 1 if the sample is an outlier and 0 if it is not.

OxCal v4.1 provides the tools to set up such models. The tools are generic and allow a wide variety of models to be employed. However it should be stressed that it is usually best to keep things fairly simple and in most cases one model should be all that is required. The two commands that have been introduced to provide outlier analysis are:

`Outlier_Model(`[*name*`,`] *distribution* [`,`*scale* [`,`*type*]] `);`
`Outlier(`[*name*`,`] [*prior*] `);`

The `Outlier_Model()` command defines the model and the `Outlier()` command allows it to be applied to specific radiocarbon dates or other likelihood information in the model. The parameters of the model are:

- *name* - this is the name of the model; this can be used to allow the specification of more than one outlier model; if the name is not specified, the last model defined will be used for any outlier analysis; if for example you wish to use a special model for all charcoal samples the name `"charcoal"` can be given to both the `Outlier_Model()` and the associated `Outlier()` commands.
- *distribution* - this defines how the outliers are to be distributed (distribution D_1); examples of useful distributions are `T(5)`, a student-t distribution with five degrees of freedom, `N(0,1)`, a simple normal distribution, or `Exp(1,-10,0)`, an exponential distribution with an exponential constant τ of 1 taken over the range -10 to 0.

- *scale* - this defines the scaling of the outliers, expressed in powers of 10; this can be a single number such as `0` for no scaling or `2` for a scale of 100 years; it can also be a distribution (distribution D_2) such as `U(0,4)` for a scale of anywhere between 1-10000 years - in this case the analysis will determine the appropriate scale.
- *type* - this defines the kind of outlier you have; the options are `"t"` for outliers in the time variable, `"r"` for those in the radiocarbon isotope ratio and `"s"` for those that scale with the uncertainty in the radiocarbon concentration.
- *prior* - for any specific measurement this defines the prior probability that the sample is an outlier; a typical value for this would be `0.05` for a 1 in 20 chance that the measurement needs to be shifted in some way. The posterior probability for the measurement being an outlier will be determined by the analysis.

The *distribution* and *scale* parameter can be defined in a number of ways in OxCal. They can either be constant numbers (this only makes sense for the scale) or they can be distributions. The relevant distributions defined in OxCal are shown in Table 1.

To see how these commands are to be used in practice we will look at some specific applications. You can also see the examples in the following section.

First of all we will consider the situation where the radiocarbon measurement itself might be at fault. We will further assume that any offsets are in proportion to the uncertainty quoted in the date. In this situation the model outlined by Christen (2003) is most appropriate. In this model (s-type) any shift in the measurement is drawn from a normal distribution which has double the uncertainty of the measurement:

$$
\begin{aligned}
r_i &= r(t_i) + \epsilon_i + \phi_i \delta_i s_i && (3)\\
\epsilon_i &\sim N(0, s_i^2 + (s(t_i))^2) && (4)\\
\phi_i &\sim \text{Bernoulli}(q_i) && (5)\\
\delta_i &\sim N(0,2) && (6)
\end{aligned}
$$

Definition	**Example**	**Meaning**
`Exp(`[*name*`,`] *tau* `,` *from* `,` *to* [`,`*resolution*]`);`	`Exp(1,-10,0)`	exponential distribution range -10 - 0 with $\tau = 1$
`N(`[*name*`,`] *mu* `,` *sigma* [`,`*resolution*]`);`	`N(0,2)`	Normal distribution $\mu = 0$, $\sigma = 2$
`T(`[*name*`,`] *freedom* `,` [*scale* [`,`*resolution*]]`);`	`T(5)`	student t-distribution 5 degrees of freedom
`U(`[*name*`,`] *from* `,` *to* [`,`*resolution*]`);`	`U(0,1)`	uniform distribution range 0 - 1

Table 1: Distribution definitions in OxCal; the optional *resolution* parameter defines the bin size during the MCMC analysis - if not specified a suitable default is chosen

where q_i is the prior probability that the sample is an outlier. In OxCal this can be specified by:

```
Outlier_Model("SSimple",N(0,2),0,"s");
```

Effectively this model draws the shifts from a normal distribution with a mean of zero and a standard deviation of 2 and they are then multiplied by the uncertainty in the date and applied to the radiocarbon measurement. So if for example the uncertainties in all the measurements are 50, the possible shifts are drawn from a normal distribution with a mean of zero and a standard deviation of 100. This is the default model for radiocarbon dates if no other is specified.

Supposing instead we have some other reason why the radiocarbon dates and those in the calibration curve may not be the same - perhaps there is possible contamination, or addition of radiocarbon from other reservoirs. In such cases (r-type) the offsets will not be related to the uncertainty in the measurement. In these cases our outlier model is modified:

$$
\begin{aligned}
r_i &= r(t_i) + \epsilon_i + \phi_i 10^u \delta_i && (7)\\
\epsilon_i &\sim N(0, s_i^2 + (s(t_i))^2) && (8)\\
\phi_i &\sim \text{Bernoulli}(q_i) && (9)\\
\delta_i &\sim D_1 && (10)\\
u &\sim D_2 && (11)
\end{aligned}
$$

where u is a scaling parameter common to the model as a whole and where D_1 and D_2 are distributions we can choose, according to modeling preference, from those given in Table 1. We might know the scale of such offsets, in which case we can fix u. If we know that they are likely to be of the order of a hundred years we can let them be drawn from a normal distribution with a mean of zero and a standard deviation of 100, by setting $u = 0$ and D_1 to $N(0, 100)$:

```
Outlier_Model("RSimple",N(0,100),0,"r");
```

If we do not know what sort of offset we are expecting we can allow the model to find the scale (anywhere between 10^0 and 10^4) and so use $U(0, 4)$ for D_2 and let the possible shifts be drawn from a longer tailed student-t distribution by using $T(5)$ for the distribution D_2 instead:

```
Outlier_Model("RScaled",T(5),U(0,4),"r");
```

The case where there might be a systematic offset between the measurements and the calibration curve (d-type) is a special case and is discussed in section 3.4.

In many cases though, the possible offsets are not in the radiocarbon scale but in the time scale (t-type). This type of outlier is applicable to other dating methods as well as radiocarbon. In the radiocarbon case we would then define:

$$
\begin{aligned}
r_i &= r(t_i + \phi_i 10^u \delta_i) + \epsilon_i && (12)\\
\epsilon_i &\sim N(0, s_i^2 + (s(t_i + \phi_i 10^u \delta_i))^2) && (13)\\
\phi_i &\sim \text{Bernoulli}(q_i) && (14)\\
\delta_i &\sim D_1 && (15)\\
u &\sim D_2 && (16)
\end{aligned}
$$

Again we might know the time-scale. For example if we have bioturbation in a sediment it might add a temporal offset between primary deposition and final location which is of the order of a hundred years. We could express this as:

```
Outlier_Model("TSimple",N(0,100),0,"t");
```

This is the default outlier model applied for non-radiocarbon measurements if no other is specified. In most cases however we do not know the scale of any such offsets and so a more general model is more appropriate:

```
Outlier_Model("General",T(5),U(0,4),"t");
```

This is the model the author would recommend for most purposes. It draws from a long-tailed distribution (D_1 is $T(5)$) and so will not be too affected by the odd extreme outlier and the scale (determined by the analysis) can be anywhere between 10^0 and 10^4 years (D_2 is $U(0, 4)$).

In some instances we might wish to use a more specific model. For example consider the case of charcoal samples. These are often discounted or used only as a `terminus post quem`. However in reality we know rather more than this. In particular many charcoal dates are likely to be only very slightly earlier than the date of deposition with a long tail of older dates from old wood or redeposited charcoal. Such a distribution is likely to be approximately exponential (as suggested by Nicholls and Jones (2001)) but with an unknown time constant (longer than a year but shorter than a thousand years). This can all be put into an outlier model suitable for charcoal:

```
Outlier_Model("Charcoal",Exp(1,-10,0),U(0,3),"t");
```

Here we only allow outliers to be older - so the exponential distribution is taken to run from -10 to 0 with a time-constant of 1. The shifts are then scaled by a common scaling factor that can lie anywhere between 10^0 and 10^3 years. In the case of charcoal samples we know that all samples are expected to be outliers (that is all earlier than context) and so they should be given a prior outlier probability of 1.

Ref	Lab ref.	Date	±	Prior	Posterior
QS1	T18161a, aa	2818	26	0.05	0.08
QS2	RTT3932.3-6	2692	24	0.05	1.00
QS3	RTT3931.3-5	2911	26	0.05	0.62
QS4	LSC3931.1	2853	25	0.05	0.03
QS5	GrN27719	2895	25	0.05	0.33
QS6	RTT 3853.1,3,4	2753	22	0.05	1.00
QS7	T3930	2800	25	0.05	0.33
QS8	T3933a, aa	2882	28	0.05	0.10
QS9	GrA25535	2864	40	0.05	0.02
QS10	GrA25710	2818	38	0.05	0.04
QS11	GrA25768	2897	44	0.05	0.06

Table 2: Dates from Tell Qasile X; the prior probability for each measurement being an outlier has been set to 0.05 (or 5%); the analysis output provides posterior probabilities for each measurement being an outlier

3 Examples

3.1 Combination of dates for samples of the same age

One situation where outlier detection can be useful is when you have a large number of radiocarbon dates all pertaining to one context but measured by different laboratories using different techniques. The congruity of such a set can be tested using the non-Bayesian χ^2 test of Ward and Wilson (1978). However, what do you do if the test fails? Using outlier detection you can down-weight those measurements which disagree most with the others and also identify which these are. As an example we will take radiocarbon dates from the important context X in Tell Qasile as reported in Boaretto et al. (2005) and Sharon et al. (2007). The measurement history is complicated and will not be discussed here. We will take 11 of the measurements reported for this context, all of which are supposed to be the same age.

The dates are all entered with a prior probability of being an outlier of 0.05. The model outlined by Christen (2003) has been applied but the treatment of the errors in the calibration curve is slightly different (see section on mathematical details below). This set of dates fails the χ^2 test (df=10, T=70 cf. 18.3) but the advantage of this kind of analysis is that in a controversial case like this you do not need to make a qualitative assessment of which dates are most likely to be wrong. You can see from Table 2 that two of the dates are identified as being definite outliers (QS2 and QS6). One other date (QS3) is also more likely to be an outlier than not.

Of course what any statistical analysis cannot do is identify the reasons why there are outliers. It could be that some of the samples really are of a different age, that there are contaminants present in some of the samples or that there is a measurement problem of some kind. Outlier analysis is useful, however in identifying which samples are most likely to be significantly wrong and providing an objective estimate of the true age of the sample set.

```
Outlier_Model(N(0,2),0,"s");
R_Combine("")
{
 R_Date("QS1", 2818,26){Outlier(0.05);};
 R_Date("QS2", 2692,24){Outlier(0.05);};
 R_Date("QS3", 2911,26){Outlier(0.05);};
 R_Date("QS4", 2853,25){Outlier(0.05);};
 R_Date("QS5", 2895,25){Outlier(0.05);};
 R_Date("QS6", 2753,22){Outlier(0.05);};
 R_Date("QS7", 2800,25){Outlier(0.05);};
 R_Date("QS8", 2882,28){Outlier(0.05);};
 R_Date("QS9", 2864,40){Outlier(0.05);};
 R_Date("QS10",2818,38){Outlier(0.05);};
 R_Date("QS11",2897,44){Outlier(0.05);};
};
```

Figure 1: Model specification for outlier analysis of Tel Qasile dates

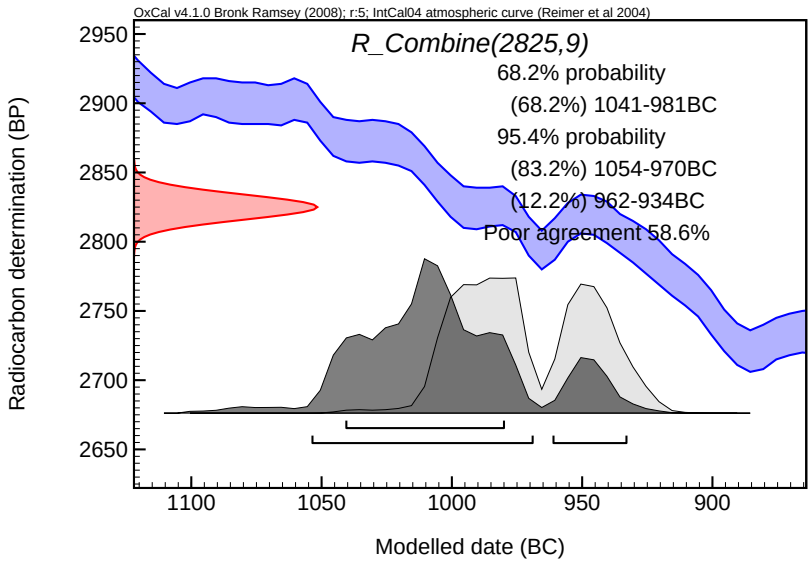

Figure 2: Combination of dates from tel Qasile X using the outlier analysis. The results of simple combination can be seen as an outline distribution in light grey; the results of the analysis are shown in darker grey and provide a significantly different age estimation.

```
Outlier_Model("General", T(5), U(0,4),"t");
P_Sequence("Simulation 4",1)
{
 Boundary(){z=480;};
 R_Date("4T47",11000,50){z=470; Outlier(0.05);};
 R_Date("4T46",11023,50){z=460; Outlier(0.05);};
 ...
 R_Date("4T2", 6901, 50){z=20;  Outlier(0.05);};
 R_Date("4T1", 6377, 50){z=10;  Outlier(0.05);};
 Boundary(){z=0;};
};
```

Figure 3: Model specification for outlier analysis of the sedimentary sequence for dataset 4 from Blockley et al. (2007)

3.2 Temporal outliers in a sedimentary sequence

The next example we will turn to is the situation where you have a sedimentary sequence where some of the samples are out of context and therefore give the wrong age for their depth. Such a situation might arise where there is significant bioturbation. To illustrate this example we can look at the simulation dataset 4 shown in Figure 6 of Blockley et al. (2007). In this case some of the datapoints had been deliberately offset from their expected values to simulate the effect of outliers. Without the use of outlier analysis it is necessary to work through the sequence eliminating those samples that have very low agreement indices in order to get a consistent model. In particular one or two of the points are so far out that the model will not run with them included.

However it is possible to use the general temporal outlier model described above instead of such a laborious and subjective procedure. Figure 3 shows how such a model is specified and Figure 4 the results of such an analysis, using a model averaging approach.

3.3 Treatment of charcoal samples

Here we give a hypothetical example to show how this might work in practice. We have a single phase of occupation which is dated by some bone dates and a series of charcoal dates. The charcoal is not short-lived and so we assume that it must always be older than its context. Figure 5 shows how such a model is specified and Figure 6 shows the results of the analysis.

In this particular case it can be deduced that the time-constant for residence of charcoal on the site lies in the range 10-100 years (see Figure 6c). In this instance the charcoal dates do add significantly to the model - many of the samples are no older than the bone dates and therefore provide important information on the date of the end of the phase and on its duration.

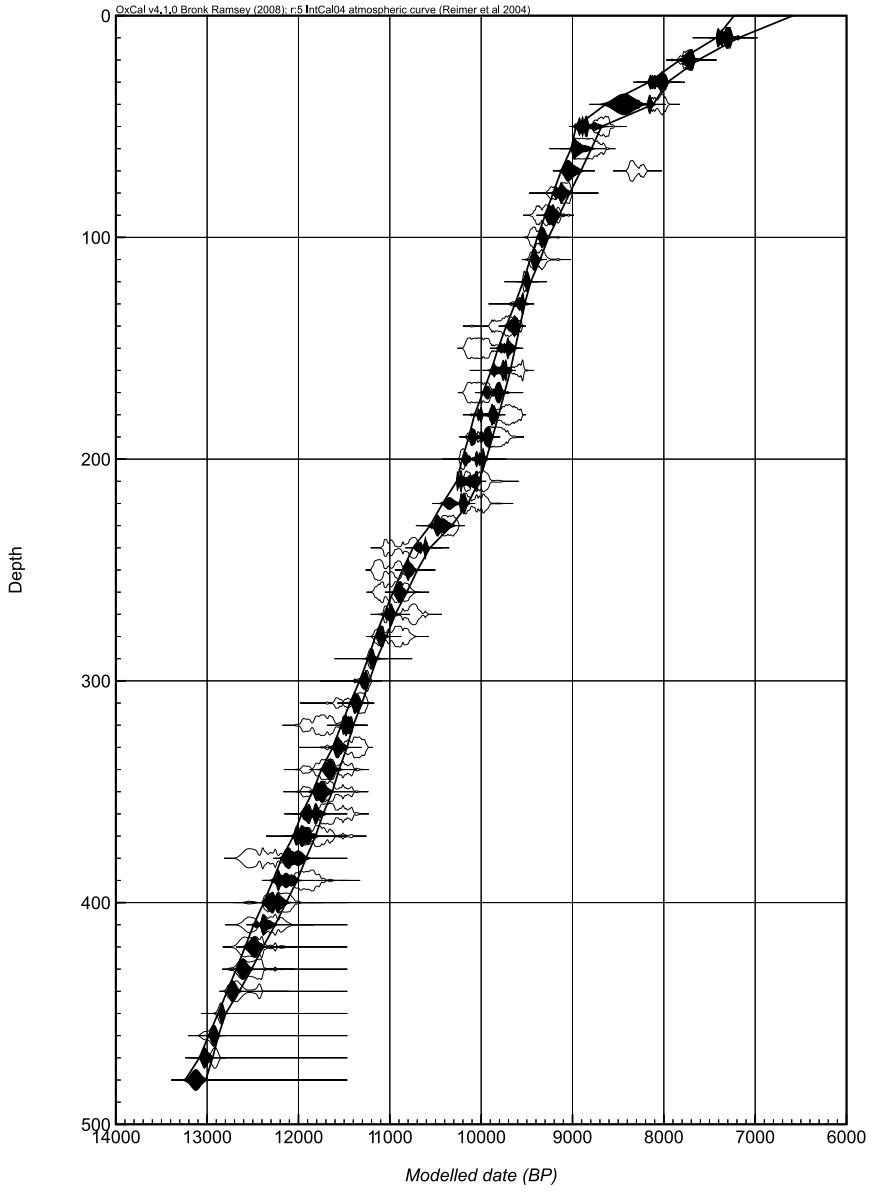

Figure 4: Age depth model for dataset 4 from Blockley et al. (2007) using outlier analysis. The results of simple calibration can be seen as an outline distribution with white fill; the results of the analysis are shown in black; you can see that some dates which are clearly outliers (such as the 7th from the top) are ignored in the analysis; this approach removes the need to weed out outliers manually before conducting such an analysis.

```
Outlier_Model("Charcoal",Exp(1,-10,0),U(0,3),"t");
Sequence()
{
 Boundary("Start 1");
 Phase("1")
 {
  Label("Bone samples");
  R_Date(1000, 20);
  R_Date(1060, 20);
  R_Date(1020, 20);
  R_Date(1070, 20);
  Label("Charcoal samples");
  R_Date(1200, 20){Outlier(1);};
  R_Date(1000, 20){Outlier(1);};
  R_Date(1030, 20){Outlier(1);};
  R_Date(1010, 20){Outlier(1);};
  R_Date(1070, 20){Outlier(1);};
  R_Date(1050, 20){Outlier(1);};
  R_Date(1130, 20){Outlier(1);};
  R_Date(1070, 20){Outlier(1);};
  R_Date(1100, 20){Outlier(1);};
 };
 Boundary("End 1");
};
```

Figure 5: Model specification for a phase with bone and charcoal dates. Note that the `Outlier` command links to the last specified `Outlier_Model` (in this case `"Charcoal"` if no *name* is specified. The command `Outlier("Charcoal",1)` could be used in each case instead.

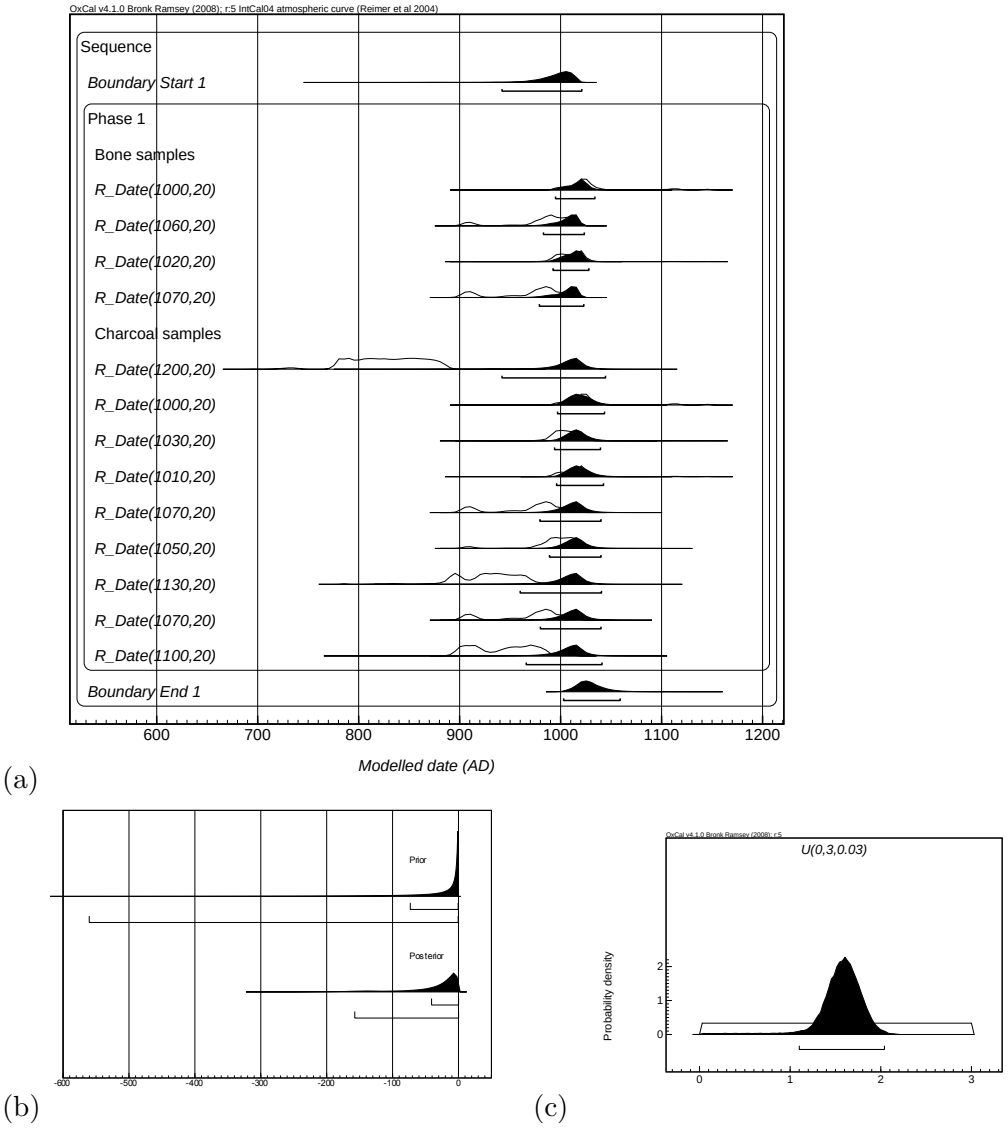

Figure 6: The results of the analysis of the model with charcoal dates. The charcoal dates act as a *terminus ante quem* for the end of the phase but some are clearly much earlier than the start as you can see in the upper plot (a). The outline distributions show the simple calibrations and the black distributions show the estimate of the deposition dates of the samples (t_i). In the lower left plot (b) you can see both the effective prior (from equation 88) and the posterior distribution of the outlier offsets ($10^u \delta_i$) which give an estimate for the charcoal ages on the site. In the lower right plot (c) the estimated time-scale (in powers of 10) for charcoal residuality on the site (the posterior distribution for u with the uniform prior shown in outline).

3.4 Systematic offsets relative to the calibration curve

As an example of such an application we can consider the dataset of Imamura et al. (2007). They give data for two tree ring sequences from Japan, one of which is a reference data-set of known age (outer ring 350 AD) and another sample data-set from a short 43 sequence which has been dendrochronogically dated (outer ring 389 AD). The standard D_Sequence analysis can be carried out for both series but, although the reference series correctly dates to 335-357AD (95.4% probability), the other sample has a bi-modal distribution with ranges 295-309 AD (86%) or 373-387 AD (9%) just missing the true value. If we now reanalyse the same two series together with a systematic ΔR value of 0 ± 10 the first still dates to a range 334-361 AD (95.4%) consistent with the correct value. The other still has a bi-modal distribution but the ranges are now 295-317AD (47.5%) and 370-395AD (47.9%) which is now in good agreement with the true value. The comparison between the two analyses can be seen in Figure 7. The reason that this works is that the analysis is able to make use of the fact that both series fit to the calibration curve better with a small systematic offset relative to the calibration curve. We can also get information on the nature of this shift. Figure 7 shows the prior and posterior for the ΔR, showing, in the posterior, a bi-modal distribution. The shift to positive ΔR gives the 'correct' fit whereas that to the left gives an equally good but 'wrong' fit for the data.

The analysis works well in this case because the reference data-set is effectively able to inform the model about the offset - even though we have not used the calendar age in the analysis. If we use the same ΔR value of 0 ± 10 and analyse the problematic sample series on its own we still do better than with no allowance for ΔR with a bi-modal range of 294-325 AD (83.3%) or 374-391 AD (12.1%) which is just in agreement with the true value. There is still a substantially higher earlier peak since, even allowing for offsets, this series does match the earlier part of the curve slightly better (see Figure 7). What is clear from this analysis is that even a small allowance for systematic offsets can have a significant effect on the accuracy of the result. What the statistical analysis cannot tell us is whether the offset is due to differences in measurement or a true regional offset.

This kind of robustness test is very important even in cases where we do not expect outliers.

4 Statistical details

In general the treatment of outliers described here is embedded in more general Bayesian analysis. Bayes theorem tells us that:

$$p(\mathbf{t}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{t})p(\mathbf{t}) \tag{17}$$

where $\mathbf{t}$ are the set of parameters and $\mathbf{y}$ the observations or measurements made. In this equation $p(\mathbf{t})$ gives our *prior* knowledge about the parameters (this can include phase models, sequences or deposition models as required). The part of the equation most important for this paper is the likelihood $p(\mathbf{y}|\mathbf{t})$ which is used to work out our *posterior* $p(\mathbf{t}|\mathbf{y})$. In many cases (when the data are conditionally independent) it is possible to factorize the likelihood into individual elements:

$$p(\mathbf{y}|\mathbf{t}) = \prod_i p(y_i|t_i) \tag{18}$$

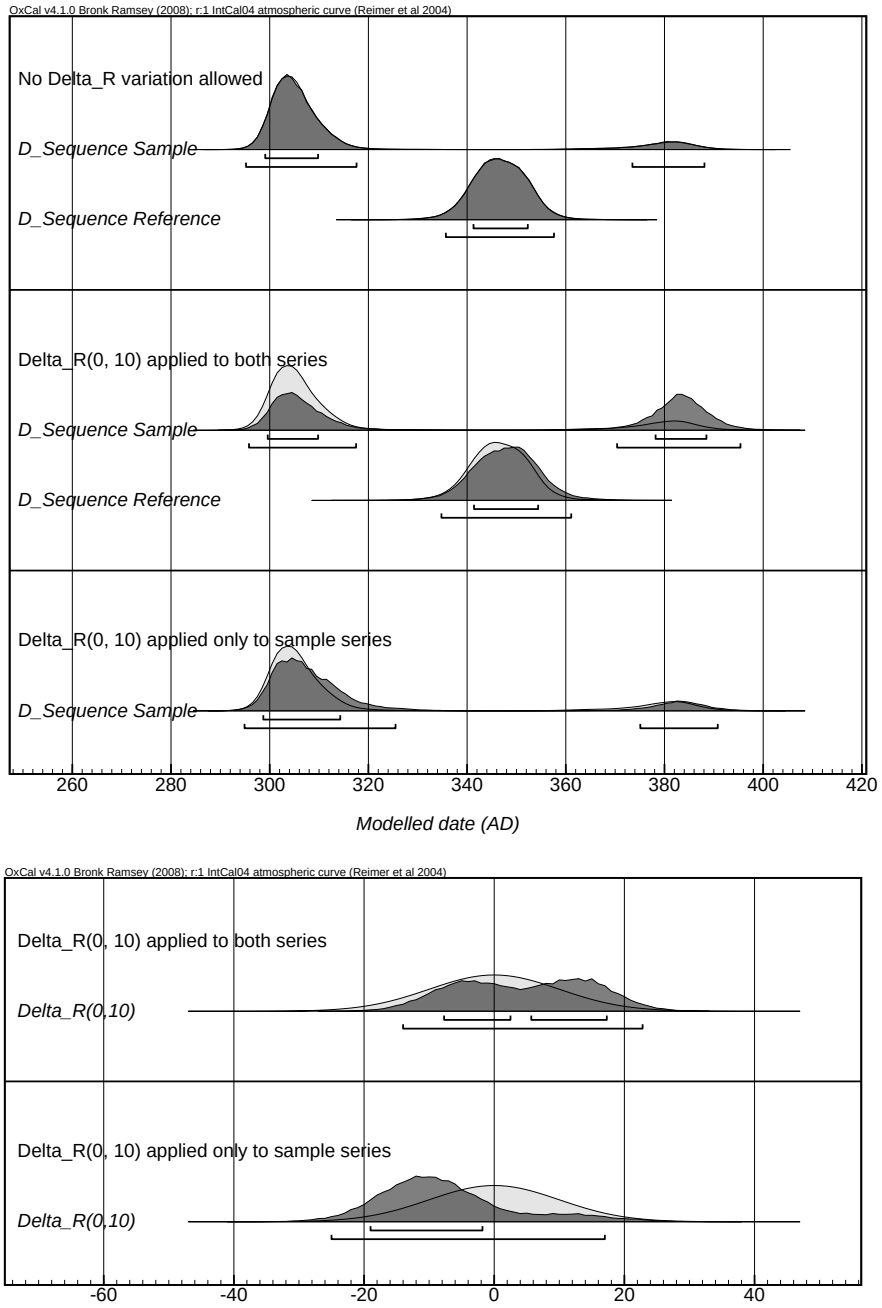

Figure 7: The results of the re-analysis of the datasets of Imamura et al. (2007). The upper figure shows the modelled end date for the sequences using the models described in the text (in dark grey) compared to those with no allowance for offsets (in light grey). The lower figure shows the posterior density estimates of the true reservoir offset (in dark grey) together with the priors (in light grey).

We use the approach of Abraham and Box (1978) as used by Christen (1994b) to deal with observations which are spurious in some way. In order to formalise this we consider the form of the likelihood function $p(y_i|t_i)$. This will in general be some function of the observed variable(s) y_i and the parameter(s) t_i involved, where the age determination are indexed $i = 1, 2, \ldots, n$:

$$p(y_i|t_i) = l_i(y_i, t_i) \tag{19}$$

In the case of radiocarbon dating the observation consists of both the radiocarbon measurement r_i and its uncertainty s_i. To use this for calibration we also need to have a calibration curve which gives the expected radiocarbon concentration, $r(t)$ and the uncertainty $s(t)$ both as a function of calendar time. Using the usual error model, with an error of ϵ_i for each measurement:

$$\begin{aligned} r_i &= r(t_i) + \epsilon_i & (20)\\ \epsilon_i &\sim N(0, s_i^2 + (s(t_i))^2) & (21) \end{aligned}$$

the radiocarbon likelihood function then becomes:

$$p(r_i|t_i) = l_i(r_i, s_i, t_i) \propto \frac{\exp\left(-\frac{(r_i - r(t_i))^2}{2(s_i^2 + (s(t_i))^2)}\right)}{\sqrt{s_i^2 + (s(t_i))^2}} \tag{22}$$

In order to model the possibility of outliers we introduce two more parameters for each observation. These parameters are ϕ_i which can take values 1 (if the observation is spurious and an outlier) or 0 (if the observation is correct) and δ_i which defines the offset in the observation if it is spurious. We define the priors for ϕ_i to be some predetermined value q_i if it is 1 (an outlier) and $(1 - q_i)$ if it is 0 (not an outlier):

$$\phi_i \sim \text{Bernoulli}(q_i) \tag{23}$$

Where the Bernoulli distribution is:

$$X \sim \text{Bernoulli}(q) \Leftrightarrow Pr(X = x) = q^x(1 - q)^{1-x} \tag{24}$$

The prior for δ_i also needs to be specified and normalised. We now need to consider the different forms of outliers.

4.1 Outliers with respect the time parameter (t-type)

In this case what we are essentially modelling for is not a wrong measurement of a variable but a wrong interpretation in terms of the parameters of the model. In this case we define the offset in terms of t_i so that in the simplest case:

$$p(y_i|t_i, \phi_i, \delta_i) = l_i(y_i, t_i + \phi_i\delta_i) \tag{25}$$

Such a model can be specified in OxCal by specifying the prior distribution for δ_i and the outlier probability q_i. For example:

```
Outlier_Model("TSimple",N(0,100),0,"t");
R_Date("OxA-12345",1423,23){Outlier(0.1,"TSimple");};
```

sets up the following priors and likelihood:

$$\begin{aligned} r_i &= r(t_i + \phi_i\delta_i) + \epsilon_i && (26)\\ \epsilon_i &\sim N(0, s_i^2 + (s(t_i + \phi_i\delta_i))^2) && (27)\\ \delta_i &\sim N(0, 100) && (28)\\ \phi_i &\sim \text{Bernoulli}(0.1) && (29)\\ p(r_i|t_i, \phi_i, \delta_i) &= l_i(r_i, s_i, t_i + \phi_i\delta_i) && (30)\\ &\propto \frac{\exp\left(-\frac{(r_i - r(t_i+\phi_i\delta_i))^2}{2(s_i^2+(s(t_i+\phi_i\delta_i))^2)}\right)}{\sqrt{s_i^2 + (s(t_i + \phi_i\delta_i))^2}} && (31)\end{aligned}$$

This is the default outlier model in OxCal for everything other than radiocarbon dates. However in many cases we would rather not specify the functional form of the prior for δ so definitely. For this reason we introduce a further model parameter, u which provides the scale for all of the outliers associated with the model. In this model the likelihood becomes:

$$p(y_i|t_i, \phi_i, \delta_i, u) = l_i(y_i, t_i + 10^u\phi_i\delta_i) \quad (32)$$

We now need to provide a prior for u as well and this is given as another parameter in the OxCal model definition. In addition it is better to use a longer tailed distribution than a normal distribution, and the student-t distribution with about 5 degrees of freedom is probably most useful for this (Venables and Ripley, 2002, p121). The reason for using such a long tailed distribution in this type of model is that, under the processes leading to temporal outliers there are sometimes a few very extreme outliers and we do not wish the modelled outlier distribution to be too heavily dependent on these. Putting all of this together, the following is a reasonable general outlier model for chronological applications:

```
Outlier_Model("TScaled",T(5),U(0,4),"t");
R_Date("OxA-12345",1423,23){Outlier("TScaled",0.1);};
```

This sets up the following priors and likelihood:

$$\begin{aligned} r_i &= r(t_i + 10^u\phi_i\delta_i) + \epsilon_i && (33)\\ \epsilon_i &\sim N(0, s_i^2 + (s(t_i + 10^u\phi_i\delta_i))^2) && (34)\\ \delta_i &\sim T(5) && (35)\\ u &\sim U(0, 4) && (36)\\ \phi_i &\sim \text{Bernoulli}(0.1) && (37)\\ p(r_i|t_i, \phi_i, \delta_i, u) &= l_i(r_i, s_i, t_i + 10^u\phi_i\delta_i) && (38)\end{aligned}$$

where the $T(\nu)$ is the student's t-distribution with ν degrees of freedom:

$$X \sim T(\nu) \Leftrightarrow Pr(X = x) \propto \left(1 + \frac{x^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)} \quad (39)$$

The reason for choosing a log-uniform distribution for the scale of offsets, is that many complex systems exhibit power-law dependency over a range of scales and the log-uniform distribution

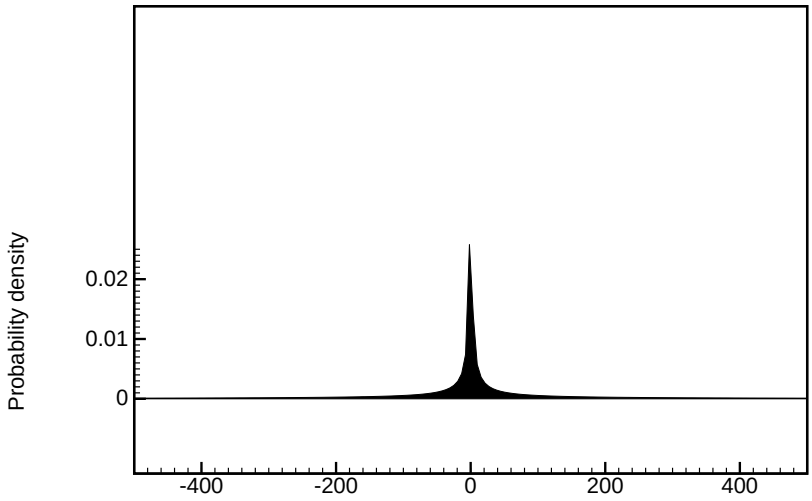

Figure 8: This shows the effective prior for the offsets in the `General` model with the full range of scales possible. The distributions is sharply peaked at zero (similar to the T(5) distribution) but with very long shallow sloped tails.

gives scale invariance. In practice the outlier scale posterior is often approximately log-normally distributed and this is easily seen as a normal distribution in u (as in Figure 6c). The effective prior for the scaled offset $10^u\delta_i$ integrated over the range of values $0 < u < 4$ and with $\delta_i \sim T(5)$ is shown in Figure 8: this is a very long tailed distribution. In most models u becomes fairly well constrained and the distribution becomes closer to the $T(5)$ distribution with an appropriate scale.

There are some dangers in having too many parameters in a model of this kind. In particular, if there are very few measurements, there may be confounding effects between the single u parameter and the δ_i parameters. For this reason with very small models it may be better to specify a fixed u. If you use the above model, you should look at the distribution for u and check that the marginal posterior is somewhat constrained (see diagnosis section below).

4.2 Outliers with underestimated uncertainties (s-type)

The construction of outliers with respect to the uncertainty quoted in the radiocarbon essentially follows the same pattern except that in this case the offset implied is in the radiocarbon measurement and not relative to the time parameter. This type of outlier obviously only makes sense in relation to radiocarbon measurements. For a simple implementation, we define the likelihood function to be:

$$p(r_i|t_i, \phi_i, \delta_i) = l_i(r_i - \phi_i\delta_i s_i, s_i, t_i) \tag{40}$$

This is essentially identical to the method proposed for generalised models in Christen (1994b) with the prior for the offset as defined in Christen (2003). The following model definition should reproduce the case where the assumed prior for $\delta_i \sim N(0, 2)$ or for the offset in the radiocarbon

date $\delta_i s_i \sim N(0, 2s_i)$:

```
Outlier_Model("SSimple",N(0,2),0,"s");
R_Date("OxA-12345",1423,23){Outlier("SSimple",0.1);};
```

and sets up the following priors and likelihood:

$$\begin{aligned} r_i &= r(t_i) + \epsilon_i + \phi_i \delta_i s_i && (41)\\ \epsilon_i &\sim N(0, s_i^2 + (s(t_i))^2) && (42)\\ \delta_i &\sim N(0, 2) && (43)\\ \phi_i &\sim \text{Bernoulli}(0.1) && (44)\\ p(r_i|t_i, \phi_i, \delta_i) &= l_i(r_i - \phi_i \delta_i s_i, s_i, t_i) && (45)\\ &\propto \frac{\exp\left(-\frac{(r_i - \phi_i \delta_i s_i - r(t_i))^2}{2(s_i^2 + (s(t_i))^2)}\right)}{\sqrt{s_i^2 + (s(t_i))^2}} && (46) \end{aligned}$$

This is the default outlier model applied by OxCal for radiocarbon dates if no other model is specified. It is equivalent to an increased variance s_i when a measurement is identified as an outlier. The program is set up to use an optional scaling parameter for this type of offset too in which case the likelihood is given by:

$$\begin{aligned} r_i &= r(t_i) + \epsilon_i + 10^u \phi_i \delta_i s_i && (47)\\ u &\sim D_2 && (48)\\ p(r_i|t_i, \phi_i, \delta_i, u) &= l_i(r_i - 10^u \phi_i \delta_i s_i, s_i, t_i) && (49) \end{aligned}$$

where D_2 is specified as for the t-type model. However, the model as defined by Christen (1994b) is reasonable for most minor measurement problems and should probably be adopted as a standard model as it is for outlier detection of this sort. This model is implemented in BCal (Buck et al., 1999), Bwigg (Christen, 2003) and Bpeat (Blaauw et al., 2003).

4.3 Outliers in radiocarbon concentration (r-type)

This implementation of the outlier model is essentially identical to that of the previous section, except that in this case we break the link with the original uncertainty in the measurement. The likelihood function is defined to be:

$$p(r_i|t_i, \phi_i, \delta_i, u) = l_i(r_i - 10^u \phi_i \delta_i, s_i, t_i) \quad (50)$$

The definition of the model is then made in exactly the same way as for the t-type outliers:

```
Outlier_Model("RScaled",T(5),U(0,4),"r");
R_Date("OxA-12345",1423,23){Outlier("RScaled",0.1);};
```

This sets up the following priors and likelihood:

$$\begin{aligned} r_i &= r(t_i) + \epsilon_i + 10^u \phi_i \delta_i && (51)\\ \epsilon_i &\sim N(0, s_i^2 + (s(t_i))^2) && (52)\\ \delta_i &\sim T(5) && (53)\\ u &\sim U(0,4) && (54)\\ \phi_i &\sim \text{Bernoulli}(0.1) && (55)\\ p(r_i|t_i,\phi_i,\delta_i,u) &= l_i(r_i - 10^u\phi_i\delta_i, s_i, t_i) && (56)\\ &\propto \frac{\exp\left(-\frac{(r_i-10^u\phi_i\delta_i-r(t_i))^2}{2(s_i^2+(s(t_i))^2)}\right)}{\sqrt{s_i^2+(s(t_i))^2}} && (57) \end{aligned}$$

4.4 Offsets relative to the calibration curve (d-type)

This type of offset is essentially the same as the r-type outlier except in this case we assume that the offset is definite (the offset probability is 1) and the same offset applies to all (or a whole set of) radiocarbon measurements. Again we introduce a new model parameter d which defines the modelled offset between the calibration curve and the set of measurements. This is a single parameter which applies to all the dates (Jones and Nicholls, 2001; Nicholls and Jones, 2001). The likelihood for each radiocarbon measurement then becomes:

$$p(r_i|t_i,d) = l_i(r_i - d, s_i, t_i) \propto \frac{\exp\left(-\frac{(r_i-(r(t_i)+d))^2}{2(s_i^2+(s(t_i))^2)}\right)}{\sqrt{s_i^2+(s(t_i))^2}} \quad (58)$$

In this equation you can see why mathematically a bias can be treated in the same way as a reservoir offset. This kind of common offset is defined as a ΔR offset. In its more usual use we have specific prior information for ΔR. However it can be used more generally and if for example there is a possible small but unknown offset between a set of measurements and the calibration curve we might set up a model with:

```
Delta_R(0,10);
R_Date("OxA-12345",1423,23);
```

This will set up the following prior and likelihood:

$$\begin{aligned} r_i &= r(t_i) + \epsilon_i + d && (59)\\ \epsilon_i &\sim N(0, s_i^2 + (s(t_i))^2) && (60)\\ d &\sim N(0, 10) && (61)\\ p(r_i|t_i,d) &= l_i(r_i - d, s_i, t_i) && (62)\\ &\propto \frac{\exp\left(-\frac{(r_i-d-r(t_i))^2}{2(s_i^2+(s(t_i))^2)}\right)}{\sqrt{s_i^2+(s(t_i))^2}} && (63) \end{aligned}$$

Several `Delta_R` statements can be used in the same model but the implementation in OxCal does not allow more than one such offset to apply to the same date.

4.5 Sampling and conditional probabilities

In all cases OxCal uses a straightforward Metropolis Hastings MCMC algorithm so only relative probabilities are important. Each parameter of the outlier model is updated individually by sampling from the full conditionals. For all of the models outlined here these are given by:

$$
\begin{aligned}
p(t_i|\mathbf{t}_{-i}, y_i, \phi_i, \delta_i, u, d) &\propto p(\mathbf{t})p(y_i|t_i, \phi_i, \delta_i, u, d) && (64)\\
p(\phi_i|t_i, y_i, \delta_i, u, d) &\propto q_i^{\phi_i}(1-q_i)^{1-\phi_i}p(y_i|t_i, \phi_i, \delta_i, u, d) && (65)\\
p(\delta_i|t_i, y_i, \phi_i, u, d) &\propto p(\delta_i)p(y_i|t_i, \phi_i, \delta_i, u, d) && (66)\\
p(u|\mathbf{t}, \mathbf{y}, \mathbf{o}, \boldsymbol{\delta}, d) &\propto p(u)\prod_i p(y_i|t_i, \phi_i, \delta_i, u, d) && (67)\\
p(d|\mathbf{t}, \mathbf{y}, \mathbf{o}, \boldsymbol{\delta}, u) &\propto p(d)\prod_i p(y_i|t_i, \phi_i, \delta_i, u, d) && (68)
\end{aligned}
$$

4.6 Radiocarbon dates all pertaining to one event (s-type or r-type)

As identified by Christen (1994b), the special case of combinations of radiocarbon dates all pertaining to one event needs to be treated slightly differently. The treatment presented there does not account for errors in the calibration curve and so a full treatment including these will be presented here.

Combination of radiocarbon dates is a two stage process. The assumption is that all of the measurements relate to one calendar time and therefore all should correspond to the same original radiocarbon concentration which we introduce as a parameter of the model ρ_c. Each measurement $(r_i \pm s_i)$ provides a likelihood function for this parameter:

$$p(r_i|\rho_c) = \frac{1}{s_i\sqrt{2\pi}}\exp(-(r_i-\rho_c)^2/(2s_i^2)) \quad (69)$$

and thus for all of the measurements:

$$p(\mathbf{r}|\rho_c) \propto \prod_i \exp(-(r_i-\rho_c)^2/(2s_i^2))/s_i \quad (70)$$

Now for convenience we define:

$$
\begin{aligned}
r_c &= \left(\sum_i r_i/s_i^2\right) / \left(\sum 1/s_i^2\right) && (71)\\
s_c &= \left(\sum_i 1/s_i^2\right)^{-1/2} && (72)
\end{aligned}
$$

$$T = \sum_i (r_i - r_c)^2/s_i^2 \quad (73)$$

$$S = \prod_i \frac{1}{s_i} \quad (74)$$

where $r_c \pm s_c$ is just the usual error weighted combination of the radiocarbon dates (Ward and Wilson, 1978). This allows us to factorise the likelihood as:

$$p(\mathbf{r}|\rho_c) \propto S \exp(-T/2)\exp(-(r_c - \rho_c)^2/(2s_c^2)) \quad (75)$$

This is just like the normally distributed likelihood for the radiocarbon ratio ρ_c you would get for a single measurement but with a mean of r_c and a standard deviation of s_c. The uncertainty in the calibration curve does not come in to the equation yet.

Now we need to consider the prior for the parameter ρ_c in the model. As for a single calibration this is given by:

$$p(\rho_c, t_c) \propto \frac{1}{s(t)}\exp(-(\rho_c - r(t_c))^2/(2(s(t))^2)) \quad (76)$$

since if we integrate over ρ_c we get a constant value, independent of t. So given this we can now integrate out the parameter ρ_c which we do not need:

$$p(\mathbf{r}|\rho_c, t_c) \propto \frac{S}{s(t)}\exp(-T/2)\exp(-(r_c - \rho_c)^2/(2s_c^2))\exp(-(\rho_c - r(t_c))^2/(2(s(t))^2)) \quad (77)$$

$$p(\mathbf{r}|t_c) \propto \int_{\rho_c=-\infty}^{\infty} p(\mathbf{r}|\rho_c, t_c)d\rho_c \quad (78)$$

$$\propto S \exp(-T/2)\frac{\exp\left(-\frac{(r_c - r(t_c))^2}{2(s_c^2 + (s(t_c))^2)}\right)}{\sqrt{s_c^2 + (s(t_c))^2}} \quad (79)$$

For the case where outliers are not considered T is a constant, and in any case S is a constant.

There are a number of useful elements that emerge from this. First of all T as defined in equation 73 is the test statistic described in Ward and Wilson (1978) which has a χ^2 distribution with $n-1$ degrees of freedom (where n is the number of combined radiocarbon dates). You can see that this is directly related (c.f. Bronk Ramsey et al., 2001) to the probability of a particular set of radiocarbon determinations for any ρ_c:

$$p(\mathbf{r}) \propto \int_{\rho_c} \exp(-(r_c - \rho_c)^2/(2s_c^2))d\rho_c \propto \exp(-T/2) \quad (80)$$

We can also now expand this treatment to deal with outliers. We offset r_i to r'_i.

$$r'_i = \begin{cases} r_i - \phi_i\delta_i s_i & \text{for un-scaled s-type} \\ r_i - \phi_i\delta_i & \text{for un-scaled r-type} \\ r_i - 10^u\phi_i\delta_i s_i & \text{for scaled r-type} \end{cases} \quad (81)$$

We repeat the same calculations to obtain r'_c (this must be repeated for each iteration of the model).

$$p(t_c|\mathbf{r}, \mathbf{o}, \boldsymbol{\delta}, d, u) \propto p(\mathbf{t})l_c(r'_c - d, s_c, t_c) \quad (82)$$

$$p(\phi_i|t_c, \mathbf{r}, \mathbf{o}_{-i}, \boldsymbol{\delta}, d, u) \propto q_i^{\phi_i}(1 - q_i)^{1-\phi_i}\exp(-(r'_i - r'_c)^2/(2s_i^2))l_c(r'_c - d, s_c, t_c) \quad (83)$$

$$p(\delta_i|t_c,\mathbf{r},\mathbf{o},\boldsymbol{\delta}_{-i},d,u) \propto p(\delta_i)\exp(-(r'_i-r'_c)^2/(2s_i^2))l_c(r'_c-d,s_c,t_c) \quad (84)$$

$$p(u|t_c,\mathbf{r},\mathbf{o},\boldsymbol{\delta},d) \propto p(u)l_c(r'_c-d,s_c,t_c)\prod_i \exp(-(r'_i-r'_c)^2/(2s_i^2)) \quad (85)$$

$$p(d|t_c,\mathbf{r},\mathbf{o},\boldsymbol{\delta},u) \propto p(d)l_c(r'_c-d,s_c,t_c) \quad (86)$$

4.7 Charcoal model

It is worth looking in a little more detail at the outlier model for charcoal samples outlined above. This model covers a range of scales u from 0 to 3 and so effective prior for a single δ would be:

$$p(\delta) \propto \int_{u=0}^{3} \frac{exp(\delta/10^u)}{10^u}du \quad (87)$$

$$\propto \frac{e^{\delta}-e^{\delta/1000}}{\delta} \quad (88)$$

where δ is only allowed to be negative. In the suggested implementation this is truncated at -10000 years and can therefore be normalised. It is a vague prior which is well behaved near $\delta = 0$ and is plotted in Figure 6b. In practice in any model the time-scale u is considerably constrained and so the distribution of outliers will be closer to a simple exponential. As above, it is important not to introduce too many different parameters into a model and in this case the model suggested is only suitable if there are many charcoal samples - if you only have one or two there will be confounding effects between the u parameter and the δ_i parameters.

5 Diagnosis and robustness testing

When using the models outlined here it is worth testing how robust the posteriors are to changes in the underlying models. This is one reason why it is important that all of the parameters of the models in OxCal are specified by the user and so can be altered to see if they affect the results. Robustness testing can easily be applied by trying different outlier models and prior outlier probabilities.

Using the scaling factor 10^u will allow the model averaging to cover a range of different scales and should achieve more robust results. This is not necessary when using the outlier methods only for outlier detection but it can be important when the method is used for obtaining realistic posterior densities from the model average. One simple example demonstrates this fairly well: we consider the case of a simple sequence of dates, with one obvious outlier, analysed under a number of different outlier models - the results of this are shown in Figure 9. Here you can see that in particular, use of the `SSimple` model (Figure 9a) actually puts quite strong constraints on measurements even if they are identified as outliers - this model is not very good for model averaging when the outliers are more extreme than the model intended. Using a longer tailed student-t distribution (Figure 9b) helps with this, but allowing the scale to adapt to the data (Figure 9c) provides a better overall average result if the scale of outliers is not known in advance.

The more complex scaled models do, however, come with the risk of confounding effects - where

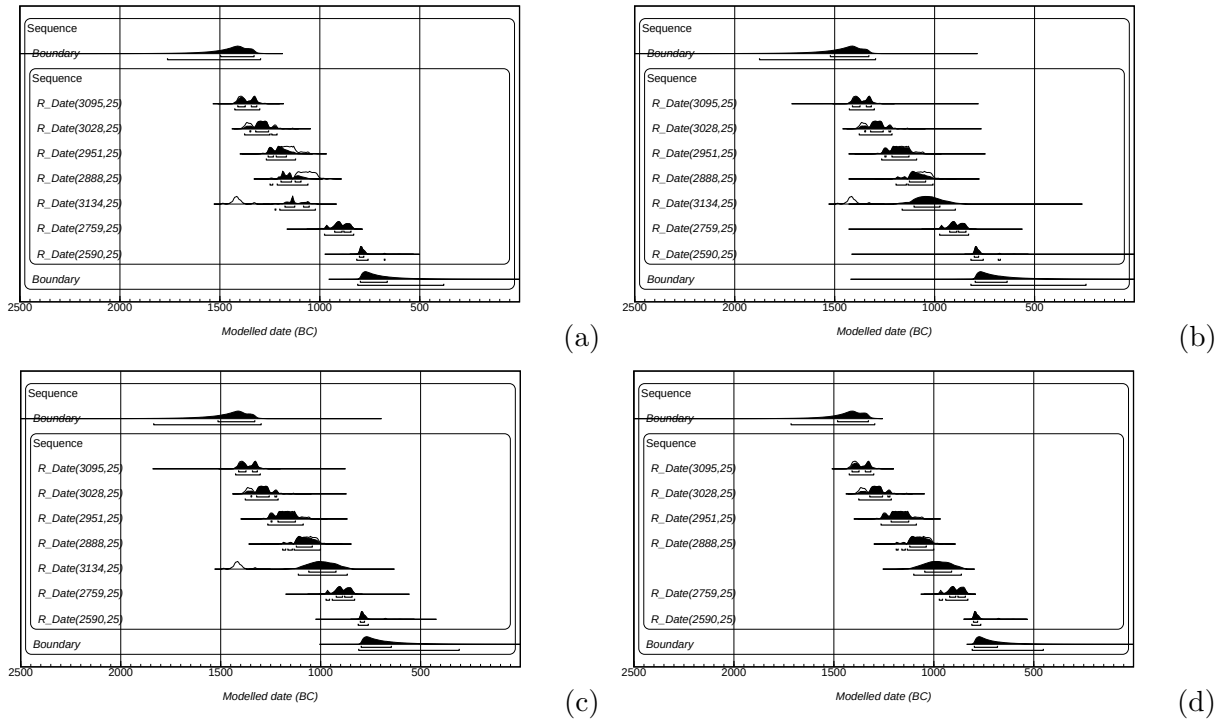

Figure 9: This show a simple sequence under a number of different outlier models: (a) uses the `SSimple` model as in Christen (2003), (b) uses a t-type outlier model with the offsets distributed as $100T(5)$ using `Outlier_Model("TFixed",T(5),2,"t")`, (c) uses the `General` model as defined above with variable scaling, (d) manually removes the obvious outlier and replaces it with an undated event. All outlier models identify the fifth radiocarbon date as a definite outlier and can be used for outlier detection. However, given that the sample is an outlier we would expect the modelled output to be similar to that shown in (d) where the date has been excluded: (a) shows that the very prescribed `SSimple` model pulls the outlier date strongly towards the measurement, even though it is an outlier, (b) which uses a longer-tailed T(5) distribution is more realistic but (c) gives a better overall model average for this situation since the dated event posteriors are very close to those shown in (d).

one parameter is played off against another and this is particularly true if for small models. It is possible for the user to check for model misbehaviour by looking at three aspects of the model output: the convergence (see Bronk Ramsey (1995)) and the posterior distribution for the scaling parameter u and the posterior outlier probabilities for ϕ_i. An number of situations can arise:

- the convergence can be very slow - this is often associated with the scale of the offsets being hard to determine; in OxCal the model may never finish running at all if a satisfactory convergence is not achieved - in such circumstances it may be necessary to use a simpler model.
- the distribution for u may be poorly constrained and extend right up to the upper limit - this is normally the consequence of using the scaled model for a dataset that is too small to support it; the results will still provide a model average over the specified scales but it would usually be better to use a simpler model in such circumstances.
- the distribution for u is well constrained at the upper limit but extends down to the lower limit - this is normally the consequence of having a dataset where there may actually be no outliers at all and you will find that the outlier posterior probabilities for ϕ_i are very close to the priors q_i; in such a situation there may be very small outliers that are undetectable and the model is reflecting this; this is not in itself a problem but it does mean that the posteriors will be at least marginally affected by the lower limit set for the scale parameter u and so a sensitivity test for this would be useful.

Ideally the scaling parameter is well defined as in Figure 6c and the convergence is reasonably rapid (though it will always be slower than for a model without outliers).

There is one other consideration which has been taken into account in the implementation of these models and this is the special case where all of the measurements are outliers. In cases where this is the intention (as in the charcoal example above) this is not a problem. However in most cases with the longer tailed distributions suggested here there is such a range of possible solutions in such cases that this can lead to extremely slow convergence. For this reason in cases where $q_i \neq 1$, and where there is more than one parameter tied to a particular outlier model, OxCal gives zero probability to the case where all measurements are outliers.

6 Conclusions

The approaches to dealing with outliers and offsets presented here are not only intended to provide detection of such offsets but also to provide good overall model averages which take into account a large range of different scenarios. For this reason it is important, as in all Bayesian analysis, that the models used reflect the underlying mechanisms. This is why a number of different models are considered in this paper:

- s-type - where the radiocarbon measurement of a particular sample is wrong for some reason: these cases can be treated with shift outliers in the radiocarbon concentration as discussed in (Christen, 1994b, 2003).

- r-type - where there are shifts between the radiocarbon concentration of the sample and its presumed reservoir but where the measurement itself is accurate: in these cases a similar approach can be used but independent of the uncertainty in the measurement itself.
- d-type - where the radiocarbon measurements are biased relative to the calibration curve - either because of problems in the measurements or because of shifts in the radiocarbon ratio of the reservoir: these cases can be effectively treated in the same way as ΔR offsets in marine radiocarbon calibration (Stuiver and Braziunas, 1993; Jones and Nicholls, 2001; Nicholls and Jones, 2001).
- t-type - where the sample measured might not relate to event being dated (t): in these cases outlier analysis using shifts in the calendar time-scale can be used.

In the case of radiocarbon samples all relating to the same event, Ward and Wilson (1978) provide a useful test of whether measurements are all compatible. However outlier analysis (s-type or r-type, depending on whether the problems are likely to relate to the measurement of the radiocarbon content of the sample) can be very useful in identifying which measurements are likely to be the outliers and giving a more objective assessment of the true age than manual rejection.

More generally where samples might or might not be outliers (as is usually the case) the methods outlined here allow an average of all of the possible combinations of rejection and acceptance of measurements to be averaged over, taking into account the posterior probabilities for such outliers. This model averaging approach is much more practical with large datasets that trying many different models each with different dates rejected. It is also more robust than selection of the outliers individually on the basis of agreement indices or outlier posterior probabilities and then just analysing one model.

The implementation of all of these techniques in OxCal v4.1 has been presented so that researchers can apply them to their own projects. The tools provided are very flexible but it should be stressed that it is probably best not to make any one model more complicated than it needs to be. For simple situations, with minor offsets, the original approach taken in Christen (2003) is likely to be sufficient for outlier detection. In larger models where displacement from context is often the main issue the general t-type model should provide a good solution. In other cases charcoal dates may need a more specific model. It is unlikely, however, that much will be gained by applying several different outlier models together, unless there are very good reasons for doing so.

The methods outlined here, if used in the right way, should start to address some of the problems associated with analysing large numbers of radiocarbon dates and help to deal with issues of over-precision which can arise if outliers and offsets are not considered at all.

7 Acknowledgements

First of all I would like to acknowledge the important work of Andres Christen who was the first to look at the Bayesian treatment of outliers in radiocarbon dates (Christen, 1994a) and who has made very valuable suggestions in the preparation of this paper. Martin Jones and Geoff Nichols

also made an important contribution in being the first to point out the need to treat ΔR offsets as a correlated parameter. Mike Dee and Richard Staff have also helped to test the implementation of the outlier models in OxCal and made many useful suggestions. Thanks are also due to Geoff Nichols who made many useful suggestions during the editing of this paper.

The research behind this paper was conducted in support of a Leverhulme funded project on 'Synchronising absolute scientific dating and the Egyptian historical chronology' (F/08 662/A). It also builds on work to develop OxCal funded by English Heritage (3164 MAIN). Without support from these institutions this work could not have been completed.

References

Abraham, B., Box, G. E. P., 1978. Linear models and spurious observations. Applied Statistics 27 (2).

Blaauw, M., Heuvelink, G. B. M., Mauquoy, D., van der Plicht, J., van Geel, B., 2003. A numerical approach to C-14 wiggle-match dating of organic deposits: best fits and confidence intervals. Quaternary Science Reviews 22 (14).

Blockley, S., Blaauw, M., Bronk Ramsey, C., van der Plicht, J., 2007. Building and testing age models for radiocarbon dates in Lateglacial and Early Holocene sediments. Quaternary Science Reviews 26 (15-16).

Boaretto, E., Timothy, J. A. J., Ayelet, G., Ilan, S., 2005. Dating the Iron Age I/II Transition in Israel: First Intercomparison Results. Radiocarbon 47 (1), 39–55.

Bronk Ramsey, C., 1995. Radiocarbon calibration and analysis of stratigraphy: The OxCal program. Radiocarbon 37 (2).

Bronk Ramsey, C., 2001. Development of the radiocarbon calibration program OxCal. Radiocarbon 43 (2A).

Bronk Ramsey, C., 2008. Deposition models for chronological records. Quaternary Science Reviews 27 (1-2), 42–60.

Bronk Ramsey, C., van der Plicht, J., Weninger, B., 2001. 'Wiggle matching' radiocarbon dates. Radiocarbon 43 (2A).

Buck, C. E., Christen, J. A., James, G. N., 1999. BCal: an on-line Bayesian radiocarbon calibration tool. Internet Archaeology 7.

Christen, J. A., 1994a. Bayesian interpretation of radiocarbon results. Ph.D. thesis, University of Nottingham.

Christen, J. A., 1994b. Summarizing a Set of Radiocarbon Determinations - a Robust Approach. Applied Statistics-Journal of the Royal Statistical Society Series C 43 (3).

Christen, J. A., 2003. Bwigg: an internet facility for Bayesian radiocarbon wiggle matching. Internet Archaeology 7.

Imamura, M., Ozaki, H., Mitsutani, T., Niu, E., Itoh, S., 2007. Radiocarbon Wiggle-matching of Japanese Historical Materials with a Possible Systematic Age Offset. Radiocarbon 49 (2).

Jones, M., Nicholls, G., 2001. Reservoir offset models for radiocarbon calibration. Radiocarbon 43 (1).

Nicholls, G., Jones, M., 2001. Radiocarbon dating with temporal order constraints. Journal of the Royal Statistical Society Series C-Applied Statistics 50.

Sharon, I., Gilboa, A., Jull, A. J. T., Boaretto, E., 2007. Report on the first stage of the iron age dating project in Israel: Supporting a low chronology. Radiocarbon 49 (1), 1–46.

Stuiver, M., Braziunas, T. F., 1993. Modeling Atmospheric C-14 Influences and C-14 Ages of Marine Samples to 10,000 Bc. Radiocarbon 35 (1).

Venables, W. N., Ripley, B. D., 2002. Modern applied statistics, 4th Edition. Springer-Verlag, New York.

Ward, G. K., Wilson, S. R., 1978. Procedures for Comparing and Combining Radiocarbon Age-Determinations - Critique. Archaeometry 20 (FEB).