Dendrochronologia 42 (2017) 51–55

Contents lists available at ScienceDirect

Dendrochronologia

journal homepage: www.elsevier.com/locate/dendro

Technical note

# Empirically-determined statistical significance of the Baillie and Pilcher (1973) *t* statistic for British Isles oak

Anthony M. Fowler[a,*], Martin C. Bridge[b,c]

[a] School of Environment, The University of Auckland, Auckland, New Zealand
[b] University College London, Institute of Archaeology, London, UK
[c] Oxford Dendrochronology Laboratory, Oxford, UK

ARTICLE INFO

*Article history:*
Received 15 September 2016
Received in revised form 1 December 2016
Accepted 25 December 2016
Available online 30 January 2017

*Keywords:*
Dendrochronology
Oak
*t* Statistic

ABSTRACT

The "Belfast method" of statistical crossdating has been widely used in the British Isles since public domain software was released by Baillie and Pilcher (1973). Although the conceptual merits of the approach are accepted, the details of the methodology have been severely criticised, including the fact that serially correlated tree-ring time series violate a fundamental requirement for the use of Students *t* statistic as a measure of statistical significance. An unfortunate consequence of this has been that *t* values are often published without reference to the associated probability of the specific value being obtained by chance. Here we present an empirical method for determining statistical significance from analysis of many misaligned inter-site correlations amongst over 2000 dated British Isles oak chronologies. Results indicate that a *t* value of 3.5 has a probability of about one in 600 for series lengths of 100+ years, but this declines (becomes less rare) as series length decreases.



## 1. Introduction

Although it has earlier antecedents, the methodological roots of modern British dendrochronology can reasonably be traced to research undertaken from the late 1960s at Queen's University, Belfast (Baillie, 1982; Fowler and Bridge, 2015). By the early 1970s, the core elements of the "Belfast method" of statistical crossdating were established and had been made publically available in the form of the computer program CROS (Baillie and Pilcher, 1973 – BP73 hereafter). In essence, the BP73 method involves computing high-pass-filtered series from raw ring-width data, which are then statistically compared by sliding them against each other at all possible positions satisfying a minimum number of years of overlap criterion. Pearson's product-moment correlation coefficient (*r*) is the goodness-of-fit statistic and Student's *t* statistic (Eq. (1)) provides "...a measure of the probability of *r* having arisen by chance" (BP73, p. 11). When applied, overlaps at many non-dating positions are expected to produce a distribution of low *t* values centred on about zero, whereas a good pattern match at the single correct-date overlap (if it exists) will produce a statistically-significant positive outlier *t* (Baillie, 1982; Fig. 1a). A value of 3.5 for *t* was adopted by the Belfast researchers as a useful, but somewhat arbitrary, indicator of a statistically-significant crossdating match.

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \quad (1)$$

where: r is the Pearson product-moment correlation coefficient and N is the number of data pairs.

The conceptual merits of the Belfast methodology are broadly accepted and it remains one of the most common inter-site oak crossdating methods used in the British Isles. It has also been adopted elsewhere by some British-trained dendrochronologists (e.g. Boswijk et al., 2014). However, caveats have commonly been expressed about statistical approaches aiding, not replacing, the skilled dendrochronologist (e.g. Baillie, 1982; Munro, 1984; Hillam, 1998), about potential misuse in the context of multiplicity (Orton, 1983; Wigley et al., 1987), and about limitations due to use of a single high-pass filtering method, as opposed to a flexible data-adaptive approach (Monserud, 1989; Yamaguchi, 1989). Moreover, the high-pass filtering method has been criticised as mathematically suboptimal (Wigley et al., 1987), with undesirable consequences related to how likely it is that correct matches will be clearly distinguishable, and also the actual statistical significance of obtained *t* statistics.

BP73 high-pass filtering entails dividing each ring width by the running mean of the five years it is the middle of, multiplying this by

* Corresponding author.
*E-mail address:* a.fowler@auckland.ac.nz (A.M. Fowler).

http://dx.doi.org/10.1016/j.dendro.2016.12.006
1125-7865/

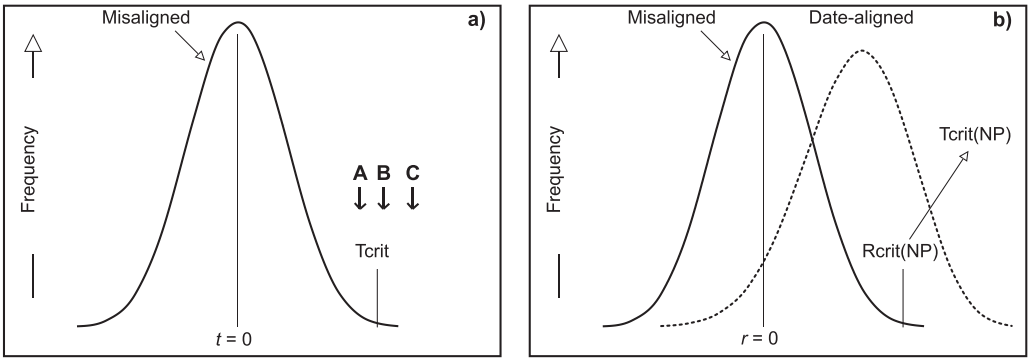


**Fig. 1.** Schematic representation of the conceptual framework. a) Standard application of the Belfast statistical crossdating method. The curve is the anticipated frequency distribution of BP73 $t$ values, $T_{crit}$ is the critical threshold required for a $t$ value to be statistically significant, and arrows denote possible date-aligned matches. b) Empirical determination of critical values of $r$ and $t$ for a given series length (N) and probability (P). The solid and dashed curves are empirically-derived frequency distributions of misaligned and date-aligned correlations for randomly-selected sample pairs. Analysis is specific to specific values of N and P, the latter derived from the relevant quantile of the misaligned correlations. $T_{crit(NP)}$ is derived from $R_{crit(NP)}$ using Eq. (1). See text for a detailed explanation of the conceptualisation.

100 to express it as a percentage, then taking the natural logs. The five-year running mean is the high-pass filter, and the log transform is intended to make the derived indices more normally distributed. An issue though is that running means cause undesirable phase distortions that "...can induce a quasi-periodic behaviour in the correlation coefficients." (Wigley et al., 1987, p. 56). The net effect is to increase the spread of the correlation frequency distribution for misaligned positions, therefore increasing the chance of obtaining a high $t$ by chance. This, in turn, means that the assumptions underpinning the use of $t$ to determine the probability associated with a given $t$ are severely compromised.

The problems associated with the BP73 high-pass filtering were recognised within a few years of the publication of CROS, and a number of alternatives and solutions were proposed, debated, and implemented in the 1980s (e.g. Munro, 1984; Monserud, 1989; Wigley et al., 1987). However, the original method continues to be widely used in the British Isles, mostly for pragmatic reasons. By the late 1980s numerous applications had shown a $t$ value of 3.5 to be "...an excellent practical choice" (Wigley et al., 1987, p. 52) and, several decades and thousands of further applications later, that probably remains a reasonable conclusion. This past use naturally creates an inertia against change, not necessarily because researchers are wedded to a methodology, but because the available database of dated sites based on that methodology is a valuable ancillary asset. Indeed, even where alternative methods are used, the Baillie and Pilcher (1973) $t$ may also be reported (Hillam, 1998).

One consequence of the recognition that the BP73 running-mean high-pass filter is suboptimal has been the divorcing of derived $t$ statistics from statements about probability. In their dating reports, practicing contract dendrochronologists tend to cite $t$ values without stating what this means in probability terms, and even peer-reviewed papers may leave the reader guessing. For example, Hillam and Tyers (1995) note the use of $t = 5$ in their reanalysis of the late John Fletcher's oak panel chronologies, but it is left to the reader to interpret how stringent a criteria that threshold actually is. Similarly, the Bridge (2012) oak dendroprovenancing paper used $t = 3.5$ as a "significant match" threshold, but again the non-specialist reader would be at a loss concerning the associated probability. Moreover, in both cases, the reader would be misinformed if they referred to standard statistical $t$-tables for guidance. Clearly, this "informal" treatment of $t$ statistics is unfortunate, but it is also a logical consequence of persisting with a methodology which is "broken" in terms of the relationship between $t$ and probability. Our aim here is to provide a solution to this issue by empirically determining the statistical significance of reported $t$ values derived using the Baillie and Pilcher (1973) method. Doing so should add value to both past and future applications of the method.

## 2. Data & methods

The oak database used in this paper is essentially the same as that compiled by Fowler and Bridge (2015), supplemented by a few additional sites. The data consist of site chronologies from the British Isles, excluding inner-London and sites with fewer than three timbers (see Fowler and Bridge, 2015 for details). A few sites subsequently found not to match the original criteria have been removed, whilst many more have been added, particularly from continuing work in North Wales, but also elsewhere and covering medieval and post-medieval periods. These changes increased the database to 2024 sites with ring-width data within the 1000–2010 CE time period considered here. Sample depth peaks at over 900 sites in the mid-1400s.

Fig. 1 schematically illustrates the conceptual framework underpinning how the British Isles oak database can be "mined" to empirically-derive probability estimates for BP73 $t$ statistics. The left panel shows a hypothetical standard application of the BP73 method, where an undated site is compared to a much longer dated series. At multiple misaligned positions, $t$ values are expected to be mostly low, centred near $t = 0$, and fit a near-normal frequency distribution (smooth curve). "$T_{crit}$" denotes a critical $t$ threshold on the right-tail of the distribution (e.g. the BP73 $t = 3.5$), above which a misaligned $t$ value is expected to be very rare, and the three arrows indicate plausible cases for the highest $t$ values for the undated series. In Case A, $t$ is less than $T_{crit}$, so would be considered to be statistically unexceptional and, normally, not worth pursuing. Cases B and C both exceed the critical threshold, so warrant further consideration (e.g. comparison of series plots). The dating confidence associated with C would normally be higher than for B, although occasionally circumstances arise where a lower value is preferred. For example, where the one or two highest $t$ values are for matches with distant sites, but there are multiple slightly lower values for sites geographically located in the region of the sample, comparisons of the time series plots may result in the lower values being accepted as the correct matching position.

A problem with the above standard practice is that probability cannot be confidently attached to $T_{crit}$ or to best-match $t$ values, because we do not know the true characteristics of the frequency distribution curve for misaligned $t$ values – especially at the extreme right-tail which describes the spurious rare high misaligned $t$ values that determine $T_{crit}$. Our contention here is that no statistical assumptions need to be made about the form

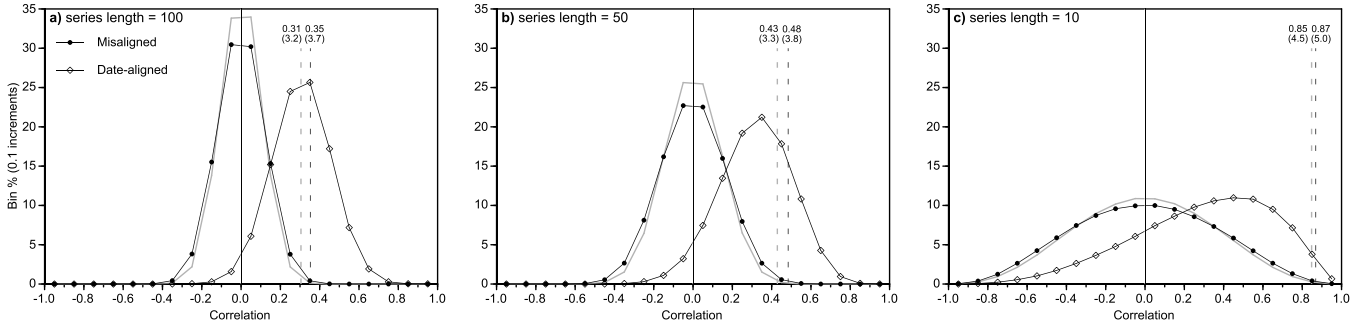


**Fig. 2.** All-site bootstrap (N = 1,000,000) correlation frequency distributions for the British Isles oak database for BP73 indices at date-aligned and misaligned positions. Series length is fixed at: a) 100 years; b) 50 years; c) 10 years. Grey curves are equivalent bootstrapped distributions for series of normally distributed random numbers. Vertical dashed lines are empirically-derived 0.999 quantile (P = 0.001) correlations for these distributions (values are above each line, with *t* from Eq. (1) in brackets).

of this distribution. Instead $T_{crit}$ can be defined empirically by randomly sampling the millions of misaligned positions in the British Isles oak database. Fig. 1b shows this schematically, but note that here it is the distribution of *r* that is presented, not *t* (a convenient simplifying step). The solid curve in Fig. 1b represents the frequency distribution for *r*, calculated for randomly-selected misaligned dates for randomly-selected site pairs, for a fixed series length (N). As with *t* in Fig. 1a, we would expect these to be centred near *r* = 0 and to be near-normally distributed, although the methodology is not sensitive to the level, spread, or shape of the distribution. If our random sample is large we can derive $R_{crit(NP)}$ for different probability thresholds from the quantiles of the random sample. For example, for a random sample of one million the 0.999 quantile (p = 0.001) will be the value equalled or exceeded by 1000 correlations. The analysis would be repeated for multiple series lengths, then the results transformed into corresponding $T_{crit(NP)}$ values. Fig. 1b (dashed line) also shows the equivalent correlation frequency distribution for date-aligned positions. This will plot to the right and may or may not have a similar shape to that for misaligned dates. Negative date-aligned correlations are possible, in part because of the wide geographical spread of the database, and it is possible that most values may fall below $R_{crit(NP)}$.

Fig. 1b is the essence of the methodology used here. For multiple series lengths (10–200 in steps of 10) we randomly sampled the British Isles oak database one million times, each time randomly selecting N years of misaligned BP73 indices from two sites. These were correlated and multiple quantiles calculated, corresponding to different probabilities (P). The net result is a two-way matrix of $R_{crit(NP)}$, showing the dependence of $R_{crit(NP)}$ on N and P. Corresponding $T_{crit(NP)}$ values are calculated from $R_{crit(NP)}$ (Eq. (1)). Note that this is more sophisticated than the approach shown in Fig. 1a, because it explicitly recognises that $T_{crit}$ is not independent of series length.

An advantage of the empirical approach is that it circumvents the need for statistical assumptions about the form of the distribution of correlation coefficients. Although we show zero-centred near-normal distributions in Fig. 1, neither is a precondition, because quantile calculation is based on ranks. Conversely, however, a problem with an empirical approach, especially when analysing the tails of distributions, is that results are likely to be sensitive to data errors. Specifically, if a site is wrongly dated, then the actual true-match correlation (presuming it exists) may be randomly selected as a misaligned correlation. Because this correlation will be biased high, relative to other misaligned correlations, it may incorrectly extend the tail of the misaligned frequency distribution and inflate $R_{crit(NP)}$ and $T_{crit(NP)}$. To mitigate this a pre-processing data step flagged all sites where the date-aligned *t* was less than 3.5 against an all-site master chronology, or where it was not the highest. The master was built by calculating the robust mean (Mosteller and Tukey, 1977) of site BP73 indices. Each of the 120 sites flagged was then reviewed to verify that the dating was not obviously suspect. This resulted in the identification of six instances where typographical errors had occurred either in the header (date) information, or the actual ring width data in transcribing files between one resource and another. The remaining 114 identified sites gave possible other matches that were dismissed as extremely unlikely (e.g. a medieval building giving a result in the Saxon period), or where multiple matches with local sites at one position were preferred over a very marginally higher *t* value with the overall mean chronology.

Although crossdating against individual site chronologies is standard practice in the British Isles, dating against master chronologies built by merging multiple sites is also quite common. If site-against-master correlations have a different frequency distribution for misaligned dates, then the associated $R_{crit(NP)}$ and $T_{crit(NP)}$ values will also differ. To test the sensitivity of the empirical estimates of $T_{crit(NP)}$ to the nature of the reference chronology, we undertook an equivalent site-against-master analysis and compared the two sets of results. The master chronology was the rebuilt all-site master chronology, following corrections for the six database errors noted above.

## 3. Results

The two solid curves in Fig. 2a show correlation frequency distributions (0.1 bins) for misaligned and date-aligned samples. The series length is 100 years and each curve is derived from one million random pairs. The black vertical-dashed line at *r* = 0.35 is the 0.999 quantile (P = 0.001), which is $R_{crit(100,0.001)}$. The corresponding $T_{crit(100,0.001)}$, from Eq. (1), is given in brackets (3.7). Comparing these results with those for an identical analysis of random numbers (light grey curve and dashed line in Fig. 2a) demonstrates how the auto-correlation in the BP73 indices results in a wider spread of misaligned correlations, with associated inflated $R_{crit(NP)}$ and $T_{crit(NP)}$ statistics. Fig. 2b and c shows results for identical analyses for series lengths of 50 and 10 years. As expected, the spread of all three correlation frequency distributions increases as series length decreases, with two important consequences. First, there is a corresponding increase in the $R_{crit(NP)}$ and $T_{crit(NP)}$ statistics required to achieve statistical significance. Second, the increasing overlap between the misaligned and date-aligned curves means that the proportion of true-date statistical matches above the threshold declines rapidly with reducing sample depth. The combined effect is represented by the area under the date-aligned curves lying to the right of the 0.999 quantile for misaligned dates.

The analyses shown in Fig. 2 were repeated for a matrix of probabilities and series lengths to explore the dependence of $R_{crit(NP)}$, and thence $T_{crit(NP)}$, on specific combinations. In effect the

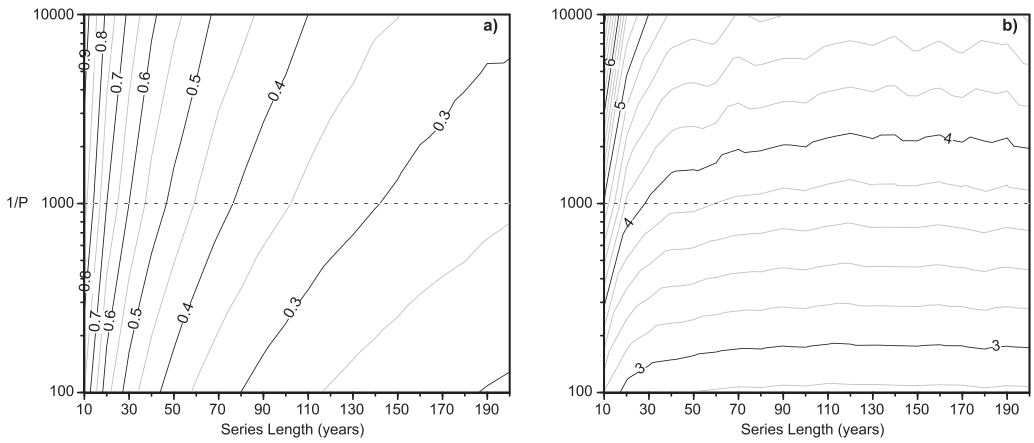


**Fig. 3.** Dependence of a) $R_{crit(NP)}$ and b) $T_{crit(NP)}$ on series length and statistical significance.

analysis needed to be repeated 20 times (once for each series length), because multiple quantiles can be calculated for each set of one million random correlations. In each case $R_{crit(NP)}$ was the statistic derived, with $T_{crit(NP)}$ obtained from Eq. (1). Fig. 3a shows $R_{crit(NP)}$, plotted as a surface against series length and 1/P, the latter restricted to what seems likely to be a useful 100–10,000 range. The $T_{crit(NP)}$ plot (Fig. 3b) demonstrates that Eq. (1) is reasonably effective at reducing sensitivity to series length, for series lengths of 40+ years. In terms of our stated aims, Fig. 3b is the most important result, providing a means to associate statistical significance with derived BP73 *t* values.

## 4. Discussion & conclusions

Although the $R_{crit(NP)}$ results (Fig. 3a) are an interim step towards $T_{crit(NP)}$ (Fig. 3b), the generic pattern apparent is of some interest. First, the near-vertical orientation of the isolines, despite the log scaling on the 1/P axis (which visually suppresses it), highlights the high sensitivity of $R_{crit(NP)}$ to series length. Second, the closer spacing of the isolines, moving from right to left across the surface, shows that sensitivity is amplified as series length declines. Third, for relatively long series, relatively low correlations are required to achieve statistical significance. All three points are consistent with expectations in terms of the statistical significance of correlation coefficients, but the empirical demonstration usefully serves to emphasise them.

Part of the point of transforming correlations into *t* values is to reduce the sensitivity of the critical threshold statistic to series length. Doing so is important in the crossdating context, because the Belfast method involves comparing multiple overlap positions with highly-variable series lengths, and derived *t* values are only meaningfully comparable if sensitivity to series-length is largely removed. Comparison of the two plots in Fig. 3 shows the extent to which the $R_{crit(NP)}$ to $T_{crit(NP)}$ transformation achieves this. The more horizontal orientation of the $T_{crit(NP)}$ isolines in Fig. 3b indicates much-reduced sensitivity to series length, especially in the case of longer series lengths (70+ years). Some residual sensitivity is apparent below about 100 years, indicating that a higher *t* value is required to achieve the same level of statistical significance as series length declines to about 70 years, and sensitivity becomes pronounced for shorter series. For example, a *t* value of 4.0 corresponds to 1/P of about 2000 for series longer than 70 years. This declines to 1000 as series length reduces to 30 years, then more rapidly to 300 at 10 years. In other words, to achieve the same statistical significance as a 100-year series with a *t* value of 4.0, a 30-year sample would need a *t* value greater than 4.2.

A *t* value of 3.5 has been a "rule-of-thumb" threshold for statistical significance for several decades, and the discussion in Baillie (1982) indicates that the value selected was intended to achieve something close to 99.9% confidence in a match at that level (i.e. one chance in 1000 that it is a spurious random correlation). The flattening of the isolines in Fig. 3b for series lengths of 70+ years indicates that a rule-of-thumb approach to $T_{crit(NP)}$ is reasonable for these longer series lengths, but caution is clearly required at shorter lengths. Moreover, the results indicate that one in 600 is closer to the mismatch chance at $t = 3.5$, for series lengths of 70+ years, dropping to one in 400 as it declines to 30 years. To achieve 99.9% confidence, a *t* value of about 3.7 is indicated, rising to about 4.0 for series lengths as short as 30 years.

All professional dendrochronologists in the British Isles impose some lower limit on the series length that they consider can reasonably be crossdated using the Belfast methodology. For example, Hillam (1998) is emphatic that samples with fewer than 50 rings should usually be rejected and that samples with fewer than 30 years should always be. Other labs are even more conservative (e.g. Belfast, Dave Brown pers com.). The results presented here indicate that there are subtle changes in behaviour as series length declines below about 70 years and that deviations become quite pronounced at less than 30 years, but they don't indicate a threshold below which statistical crossdating is invalid. For example, Fig. 3b shows that a 10-year series yielding $t = 6.0$ is statistically more significant (1/P = 3000) than a 150-year series with $t = 3.7$ (1/P = 1000). However, this does not mean that it makes sense for dendrochronologists to analyse very short series, because the chance of obtaining a significant *t* value becomes vanishingly small as the length declines (Fig. 2). For example, whereas 38% of date-aligned *t* values for series lengths of 100 years exceed the 0.999 quantile for misaligned dates, this reduces to 18% at 50 years and <2% at 10 years (i.e. a futile waste of effort).

The $T_{crit(NP)}$ surface for the sites-against-master analysis (not shown) is almost identical to the sites-against-sites surface shown in Fig. 3b. This suggests that the latter can reasonably be used to attach a probability estimate to *t* values obtained when comparing site chronologies against merged-site masters, although additional testing against regional-scale master chronologies would be useful. The important impact of crossdating against the all-site master is in fact the separation of the misaligned and date-aligned curves (Fig. 2). In essence, although $T_{crit(NP)}$ is little affected, the likelihood of a correct-date *t* exceeding it is much increased. For example, the 38% and 18% given above, for series lengths of 100 and 50 years, respectively increase to 93% and 74%.

How the $T_{crit(NP)}$ surface presented in Fig. 3b could be used in future applications, or retrospectively to add value to published

work, depends on how $t$ values have or will be used. For applications, such as the Bridge (2012) provenancing study, where a single $t$ value is calculated for each site at the dated position, application is quite simple, and essentially just removes ambiguity regarding statistical significance. For example, Fig. 3b indicates that the $t = 3.5$ used by Bridge (2012) as a metric for statistical significance corresponds to a significance level of <0.002 for series longer than about 30 years. If a similar study were to be done in future, the results presented here would allow a fixed probability to be prescribed, with the $T_{crit(NP)}$ surface interrogated to identify the series-length-dependent threshold $t$ value for each site. However, for crossdating applications where $t$ values are calculated at many overlap positions, and often for multiple sites, the single-comparison probability shown in Fig. 3b would need to be adjusted to account for multiplicity (Orton, 1983). Clearly this can only be done for previously published work if details concerning the number of overlap positions tested are given.

The results presented here are specific to British Isles oak and to BP73 high-pass filtering. Applicability to continental European oak seems likely, presuming that the autocorrelation characteristics are the same, but this would need to be verified. This is because autocorrelation is likely conjointly influenced by both biological and climatic controls. Moreover, the wealth of oak data available for continental European means that there is no reason to make assumptions about the transferability of results (i.e. Fig. 3) – because it would be a simple task to replicate the empirical methodology presented here. For other species and for alternative high-pass filtering methods assumptions of similarity are inappropriate. However, we note that the methodology presented here is generic and widely applicable. There are essentially only two prerequisites: a numerical measure of the goodness-of-fit of two time series and a sufficiently large data base to empirically determine critical thresholds for that goodness-of-fit statistic.

## Acknowledgements

We gratefully acknowledge the following individuals and organisations who contributed oak archaeological data. Data were sourced from individual sites worked on by MCB and his colleagues in the Oxford Dendrochronology Laboratory and its predecessors, especially Dan Miles, Historic England (formerly English Heritage) funded sites (where the information is in the public domain in the reports that they produce) and other site chronologies kindly supplied by Ian Tyers, the Nottingham Tree Ring Laboratory (mostly through Robert Howard) and the former Sheffield Dendrochronology Laboratory (mostly through CathyTyers), along with Irish material supplied by David Brown (Queen's University Belfast), and other sites from Anne Crone, Coralie Mills, and Rob Wilson. In addition site information was obtained from the International Tree Ring Databank, contributed by Jennifer Hillam, Tom Melvin and Keith Briffa.

## References

Baillie, M.G.L., Pilcher, J.R., 1973. A simple crossdating program for tree-ring research. Tree-Ring Bull. 33, 7–14.
Baillie, M.G.L., 1982. Tree Rings and Archaeology. University of Chicago Press, Chicago.
Boswijk, G., Fowler, A.M., Palmer, J.G., Fenwick, P., Hogg, A., Lorrey, A., Wunder, J., 2014. The late Holocene kauri chronology: assessing the potential of a 4500-year record for palaeoclimate reconstruction. Quat. Sci. Rev. 90, 128–142.
Bridge, M., 2012. Locating the origins of wood resources: a review of dendroprovenancing. J. Archaeol. Sci. 39, 2828–2834.
Fowler, A.M., Bridge, M.C., 2015. Mining the British Isles oak tree-ring data set. Part A: rationale, data, software, and proof of concept. Dendrochronologia 35, 24–33.
Hillam, J., Tyers, I., 1995. Reliability and repeatability in dendrochronological analysis: tests using the Fletcher archive of panel-painting data. Archaeometry 37, 395–405.
Hillam, J., 1998. Dendrochronology: Guidelines on Producing and Interpreting Dendrochronological Dates. English Heritage, London.
Monserud, R.A., 1989. Comments on "Cross-dating methods dendrochronology" by Wigley et al. J. Archaeol. Sci. 16, 221–222.
Munro, M.A.R., 1984. An improved algorithm for crossdating tree-ring series. Tree-Ring Bull. 44.
Mosteller, F., Tukey, J.W., 1977. Data Analysis and Regression: A Second Course in Statistics. Addison- Wesley, Reading, Massachusetts.
Orton, C.R., 1983. The use of Student's $t$-test for matching tree-ring patterns. Bull. Inst. Archaeol. Univ. Lond. 20, 101–105.
Wigley, T.M.L., Jones, P.D., Briffa, K.R., 1987. Cross-dating methods in dendrochronology. J. Archaeol. Sci. 14, 51–64.
Yamaguchi, D.K., 1989. Comments on "Cross-dating methods dendrochronology" by Wigley et al. J. Archaeol. Sci. 16, 222–224.