David G. Smith, 2020. Misplaced confidence: limits to statistical inference in cyclostratigraphy. Boletín Geológico y Minero, 131 (2): 291-307
ISSN: 0366-0176
DOI: 10.21701/bolgeomin.131.2.005

Misplaced confidence: limits to statistical inference in cyclostratigraphy

David G. Smith

15 Stratton Terrace, Truro, Cornwall, U.K.
d.g.smith@talktalk.net

ABSTRACT

Spectral (frequency-domain) analysis is used for quantitative confirmation of cyclicity in climate-proxy data. Cyclostratigraphic power spectra are typically accompanied by 'confidence limits', whether or not a statistical test has been explicitly invoked. Peaks in spectral power suggest candidate cyclic frequencies; confidence limits (CLs) appear to provide a visual guide to their relative importance, and are conventionally used in a correspondingly informal way. Confidence limits are, however, inseparable from formal tests of statistical significance; they derive from a statistical null hypothesis, and provide a threshold for its acceptance or rejection. In the procedure conventionally used in cyclostratigraphy (and implemented in several specialised software packages), noise models and confidence limits are generated automatically. Although the user may be unaware of it, the null hypothesis on which these CLs are based is calibrated for a (confirmatory) test of significance at exactly one frequency. Extending their application to an exploratory search of spectral peaks at all frequencies is statistically inadmissible. Debate over the role and correct calculation of CLs in cyclostratigraphy remains unresolved: this contribution seeks to clarify the disagreement over their use by explaining the role of CLs in statistical significance tests generally, and comparing it with their conventional use in cyclostratigraphy. Through examples of the correct and incorrect use of the conventional method, I show that the customary informal use of statistical test criteria cannot be sustained. Significance thresholds cannot be calculated in most cases; wrongly estimated confidence limits lead to false positive cycle identifications, with adverse consequences for calibration of the geological time scale.

Keywords: cyclostratigraphy, spectral analysis, confidence limits, null hypothesis.

Confianza perdida: límites a la inferencia estadística en cicloestratigrafía

RESUMEN

El anàlisis espectral (dominio de la frecuencia) es utilizado para la confirmación cuantitativa de la presencia de ciclicidad en datos climáticos o un indicador indirecto (proxy). Los espectros de potencia en cicloestratigrafía están típicamente acompañados por 'límites de confianza', independientemente de que se haya invocado o no, explícitamente, un test estadístico. Picos en el espectro de potencia sugieren frecuencias candidatas a mostrar ciclicidad; los límites de confianza (CLs) parecen proporcionar una guía visual a su importancia relativa, y son usados convencionalmente de un modo informal. Sin embargo, los límites de confianza son inseparables de test estrictos de significación estadística; derivando de una hipótesis estadística nula, y proporcionan un umbral para su aceptación o rechazo. En el procedimento que es utilizado convencionalmente en cicloestratigrafía (y que está implementeado en paquetes informáticos especializados), se generan de modo automático modelos para el ruido y los límites de confianza. Aunque el usuario puede no ser consciente de ello, la hipótesis nula sobre la que estos CLs están basados está calibrada para un test de confianza confirmatorio a exactamente una frecuencia. La extensión de su aplicación a una búsqueda exploratoria de los picos espectrales en todas las frecuencias es estadísticamente inadmisible. El debate sobre el papel y el cálculo correcto de los CLs en cicloestratigrafía no ha sido todavía resuelto: esta contribución pretende clarificar el desacuerdo sobre su uso mediante la explicación del papel de los CLs en los test de significación estadística en general, y su comparación con su uso convencional en cicloestratigrafía. A través de ejemplos sobre el uso correcto e incorrecto del método convencional, se muestra que el uso informal acostumbrado de los criterios de test estadístico no se sostienen. En la mayoría de los casos no se pueden calcular umbrales de significación; la estimación errónea de los límites de confianza conduce a la identificados de ciclos que son falsos positivos, con consecuencias adversas para la calibración de la escala de tiempo geológico.

Palabras clave: cicloestratigrafía, análisis espectral, límites de confianza, hipótesis nula.

Introduction

Power spectral analysis is much used in cyclostratigraphy, where stratification cycles are predicted to encode orbital forcing. This paper concerns the use, misuse, and non-use of the statistical test criteria that are routinely plotted with power spectra; I suggest that their validity is far more limited than is suggested by their widespread use in practice.

Walther Schwarzacher (e.g. Schwarzacher 1975, Chapter 8) was an early pioneer both of spectral analysis, and of the use of statistics to test the relative significance of power spectral peaks. Later, a key development was the introduction of a particular approach to the estimation of spectral background ('noise') by Mann and Lees (1996; ML96). Although the context of their work was climate science, their 'robust' method was judged to be appropriate for application in cyclostratigraphy.

ML96 quickly became the method of choice for power spectral analysis in cyclostratigraphy, becoming embedded in specialised software packages such as SSA-MTM (http://research.atmos.ucla.edu/tcd/ssa/) and Astrochron (Meyers, 2019a). Although other approaches continued (and continue) to be tried, ML96 remains the foundation of what I refer to here as the 'conventional' procedure in cyclostratigraphy, regarded by many as this discipline's default means of conducting spectral analysis.

Vaughan, Bailey and Smith (2011) investigated the statistical features of this conventional method, and found two important sources of false positive results, also known as Type I errors. (1) It assumes that a single class of noise model can be applied to all datasets; and (2) the widespread misunderstanding that the confidence limits it calculates can be used to test for significance at multiple spectral frequencies.

A series of critical discussion papers followed, in which these findings were applied to a number of studies, for which corrected test criteria were proposed (Smith, Bailey and Vaughan 2016; Smith and Bailey 2017a,b; Smith and Bailey 2018a, b, c; Smith, 2019). These corrections were dismissed, essentially on grounds of the irrelevance of statistical rigour to cyclostratigraphy, and confirmed a lack of interest in engaging with the statistical issues (Hinnov, Wu and Fang 2016; Andrews, Cornwell, Trewin et al. 2018; Thibault and Perdiou 2018; Hinnov, Ruhl and Hesselbo 2018; Howe, Corcoran, Longstaffe et al. 2018; Gong and Kodama 2018; Da Silva, Dekkers, De Vleeschouwer et al. 2019).

As the criticisms by Vaughan, Bailey and Smith (2011) remain technically correct, it follows that there is a major difference of opinion concerning the way that statistical criteria are applied in cyclostratigraphy. I explore this controversy by focussing on the differences between what the conventional method really does, and how it is widely applied in practice. It is important to note that it is the statistics that is the subject of my criticism, and not cyclostratigraphy in general.

This paper is organised as follows – note that the principal arguments can be followed through the figures and their captions, as well as through the text. I first (Figure 1) use a synthetic dataset to show how the conventional method identifies non-existent cycle-periods in random data. Recognising statistical multiplicity as an important source of false positives, I use dice to illustrate this concept (Figure 2). Figure 3 presents an example of the correct, single-frequency application of a conventionally calculated confidence limit using the familiar example of sunspot numbers. Figure 4 shows the conventional extension of a CL to multiple frequencies in a power spectrum of climate data, with misleading results. In Figures 5 and 6, corrections for statistical multiplicity allow both the random dataset (Figure 1) and the climate dataset (Figure 4) to be interrogated for statistical significance at multiple frequencies. Figure 7 contrasts two different analyses of a cyclostratigraphic dataset, one using the conventional – informal – approach, the other using corrected statistical criteria. I then discuss the problems raised by these examples in terms of the role of hypotheses, the useful distinction between exploratory and confirmatory modes of data analysis, and the recognition, sources, and impact of statistical multiplicity in cyclostratigraphy, before concluding with some recommendations for future practice.

Confidence limits in cyclostratigraphy

Power spectra are the standard means of evaluating the frequency-domain properties of a data-series. In cyclostratigraphy, the plots of many such power spectra also include statistical test criteria in the form of a 'noise model' and/or 'confidence limits'; these features are present mainly because the conventional procedure (and its implementation in software toolkits) integrates their calculation with that of the spectrum.

The conventional procedure thus generates a noise model/confidence limit whether or not it is the intention to conduct a statistical test; this is a key criticism in this paper. In most cases, all that is actually needed in the initial, exploratory phase of data analysis is a power spectrum; hypothesis-confirmation through statistics is rarely relevant, appropriate, or even possible at this stage of an investigation.

In statistical terms, exploratory data analysis (EDA) describes an initial phase of wide-ranging and open-ended exploration of the properties of a newly collected dataset, in a general search for patterns such as cycles. EDA leads to the erection of specific hypotheses, for statistical testing in the subsequent, confirmatory (CDA) stage of analysis, which is ideally conducted on an independent, freshly collected dataset. The pragmatic distinction between exploratory and confirmatory data analysis was introduced by Tukey (1977), and provides a useful framework for understanding the current 'crisis of confidence' in cyclostratigraphy.

A test of statistical significance (Vaughan 2013, Chapter 7) requires a null hypothesis (NH), which is usually that the data are random. If the chosen significance threshold is not reached, the NH provides the default conclusion: the data are random, and the hypothesis that a pattern (cycle) exists is not supported. Positive statistical support for the hypothesised effect requires the null hypothesis to be rejected.

For any quantitative test, the null hypothesis needs numerical expression; this is the function of the noise model, which is estimated from the power spectrum; it is usually plotted as a continuous function. Noise estimation is achieved either through some empirical best-fit procedure (Vaughan, Bailey and Smith, 2011; Weedon, this volume), or by fitting to some preferred template. The latter is the case in the Mann and Lees (1996, ML96) approach, where the noise is required to take the form of a first-order autoregressive (AR1) process.

Although usually depicted as continuous, the noise model is in fact discrete, being estimated at each of the $N/2$ frequencies at which spectral power is calculated (N is the number of points in the original data-series). Although usually represented by a single value at each frequency, the model is in fact a probability density function (PDF), such as a Gaussian distribution. The line generally used to represent the noise model (the blue line on Figure 1, for example) connects the median values of the model PDFs, of which there is one at each of the $N/2$ frequencies.

The confidence limits (CLs) also connect points on the noise PDFs at each frequency, but above the median, such that (for example) only 5% of each PDF's values lie above the 95% CL. This provides a (user-defined) threshold for testing the significance of spectral power at some frequency, and hence for either accepting or rejecting the null hypothesis at that frequency.

My objective here is to show that the confidence limits that are generated automatically by the conventional procedure are not fit for the purpose that their presence implies. In EDA/CDA terms, any confidence limit is integral to a confirmatory test of statistical significance, and has no meaning outside that context. Conventional practice, however, is to treat CLs as an exploratory tool, as a context-free guide, and as only one of several criteria available for open-ended exploration of the data.

Statistical testing is generally inapplicable to the exploratory stage of analysis, because the open-ended nature of EDA entails too many scenarios, a problem known as statistical multiplicity (discussed in more detail below). Statistical tests are appropriate in confirmatory analysis to the extent that the range of scenarios can be restricted and quantified. Conventional usage in cyclostratigraphy effectively applies confirmatory significance testing in an exploratory environment, with the unavoidable result that a high degree of statistical multiplicity downgrades the reliability of the confidence limits.

I now use examples to demonstrate how this exploratory abuse of confidence limits leads to false positive results, and how they can be corrected for their proper, confirmatory use.

The conventional procedure: false positive cycles in random data

In this paper, the 'conventional approach' to spectral analysis in cyclostratigraphy refers to the simultaneous calculation of a power spectrum, a default noise model, and one or more confidence limits (CLs), based on the method introduced by Mann and Lees (1996), and usually through use of a specialist software package. Here, I use a synthetic dataset to introduce a typical spectrogram on which spectrum, noise and CL are superimposed. Note that this example is also typical in that the noise model and CL have no explicit purpose; the reader is left to infer that the plot is for general illustration of the data in the frequency-domain. **[Figure 1]**

Figure 1 plots the conventional analysis of a 1024-point simulated data-series (Figure 1A) using a widely available software package (see figure caption for details); the procedures used for the real datasets in Figures 3 to 7 were essentially the same, with minor modifications that are explained in the relevant text and figure captions. All spectrograms in this paper are plotted against both log-log and linear-linear axes. A linear frequency axis is often preferred in cyclostratigraphic studies, but log-log plots are essential for viewing the whole frequency range, particularly when comparing the modelled spectral background with the spectrum of the data.

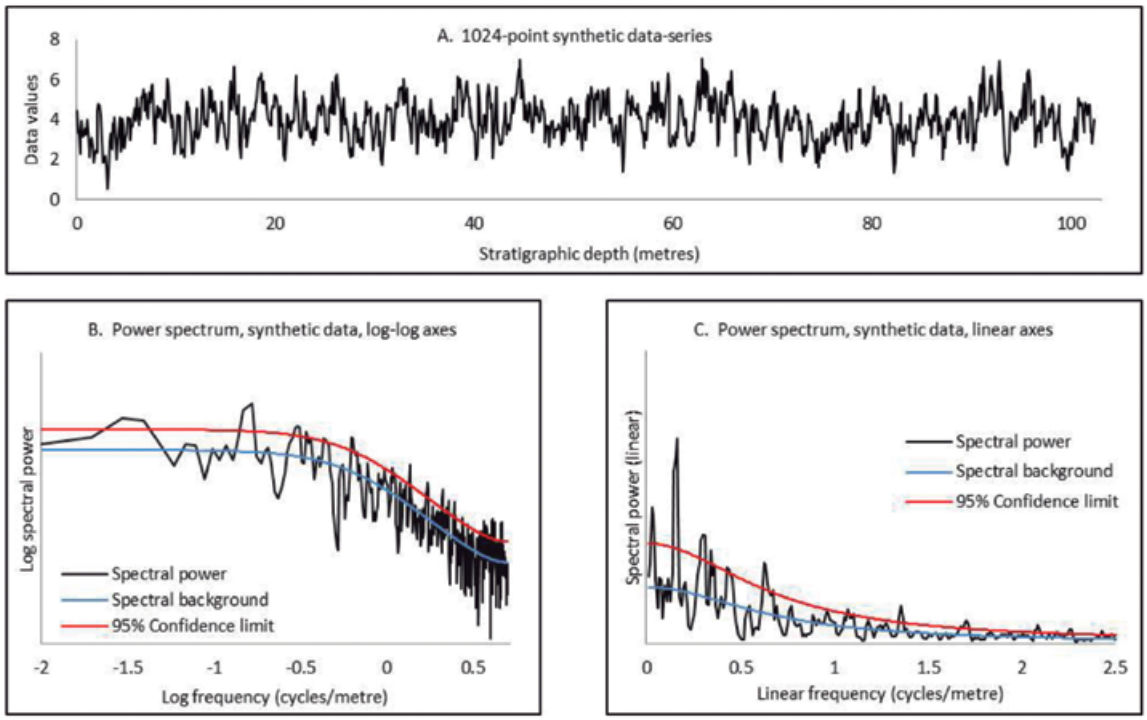

Figure 1. False positive cycles in random data: conventional spectrogram with hypothesis-free noise model and confidence limits.

Figura 1. *Falsos positivos de ciclos en datos aleatorios: espectrograma convencional con un modelo de ruido y límites de confianza sin hipótesis preconcebidas.*

An exploratory search of a conventionally calculated power spectrum finds significant cycle-periods in random data, because the confidence limit (CL) is correct only for a test at a single frequency. Chance dictates that Power>CL at a number of frequencies, just as a throw of many dice is likely to include a number of sixes (Figure 2). Figure 6 shows how to control the false positive rate for this dataset.

A: 1024-point realisation of first-order autoregressive (AR1) random process, using the ar1 function in Astrochron (Meyers, 2019a), with autocorrelation coefficient $\rho = 0.7$.

B, C: MTM spectral analysis (3 tapers), with 'robust' AR1 noise model and CL, all calculated together using Astrochron function mtmML96; results exported to Excel for plotting. B and C are identical except for their axes: log-log in B, bi-linear in C.

Colour conventions (all figures): blue=noise model (spectral background); red=uncorrected confidence limit (CL); green=corrected CL.

Figure 1B/C represents the output of the all-in-one function *mtmML96* in the Astrochron software package (Meyers, 2019a). This (and equivalent functions in other packages) applies (1) the MTM (Thomson's multi-taper) method to generate the power spectrum; and (2) the 'robust' noise modelling method of Mann and Lees (1996; ML96) to estimate the spectral background and hence one or more confidence limits. For clarity in this and other figures, I plot only a single CL.

In Figure 1B/C, the 95% confidence limit (CL) intersects the power spectrum at a number of local peaks, thus defining frequencies at which the spectrum lies above the confidence limit (Power>CL). What are the implications of this: are the data to be interpreted as cyclic at those frequencies, 'with 95% confidence'? If not, what is the function of the confidence limit? (Recall the absence of any stated objective or hypothesis.)

The dataset in this example is in fact random; it is a realisation of a simple autocorrelative process (see figure caption for details) and is therefore not periodic at any frequency. The conventional procedure has therefore generated a plot whose face-value interpretation is unambiguously misleading; the frequencies that it identifies as significant (and hence cyclic) are all false positives. The procedure's arithmetic correctly replicates the ML96 method, which is itself technically correct, yet this standardised calculation has led to a misleading juxtaposition of a confidence limit and the data's power spectrum; it has found cycles that do not exist in the data.

ML96 is primarily a method for estimating background noise from a power spectrum, a necessary step towards a statistical significance test. However, ML96 has come to be used as the default method for calculating the power spectrum, whether or not there is any intention of using statistics. In the random example above, no objective for Figure 1B/C was stated; I declared neither a scientific hypothesis nor a statistical null hypothesis. In such a lack of context, the functions of the noise model and CL are necessarily ambiguous.

Figure 2. Statistical multiplicity: many dice, many sixes.

Figura 2. *Multiplicidad estadística: muchos datos, muchos seises.*

In this analogy, the frequencies in Figure 1B/C are represented by dice; the single-test confidence limit (the red line in Figure 1) is represented by the assumption that the overall probability of seeing a six is 1:6, as it is for a single die. Multiplicity (throwing many dice) raises the probability of seeing a six; the sixes seen here are the product of chance, as are the peaks in spectral power in Figure 1.

Statistical test criteria, including confidence limits, can only be based on a null hypothesis. Given that a noise model and CLs appear in Figure 1B/C, what is the implied null hypothesis; and why are false positive cycles identified when the CL in Figure 1B/C is used to search for significant frequencies?

ML96 correctly calculates a confidence threshold for a test of spectral power. The source of ambiguity is that this test of significance is correct only if applied at exactly one frequency. Such a single-frequency test is achievable only if that frequency can be specified in advance, but this is very rarely possible in cyclostratigraphy (see Figures 3 and 4 for examples where it is possible to use ML96 uncorrected).

Thus, Figure 1 is misleading because it appears to provide a significance test that is applicable at all frequencies, yet the significance threshold (the CL) is calibrated for such a test at only one frequency. The critical, seldom acknowledged factor in this is statistical multiplicity: a concept that I now introduce through a familiar example, before applying the conventional procedure to an explicitly single-test example.

Statistical multiplicity: a simple analogy

I introduced statistical multiplicity above in the context of the distinction between exploratory and confirmatory data analysis; it is one of the two sources of false positives identified by Vaughan, Bailey and Smith (2011); and it is central to the difference of opinion over strict *versus* permissive use of confidence limits. **[Figure 2]**

Statistical multiplicity is the effect of multiple simultaneous (or repeated) applications of a significance test. Often unrecognised, it will always – if uncorrect-

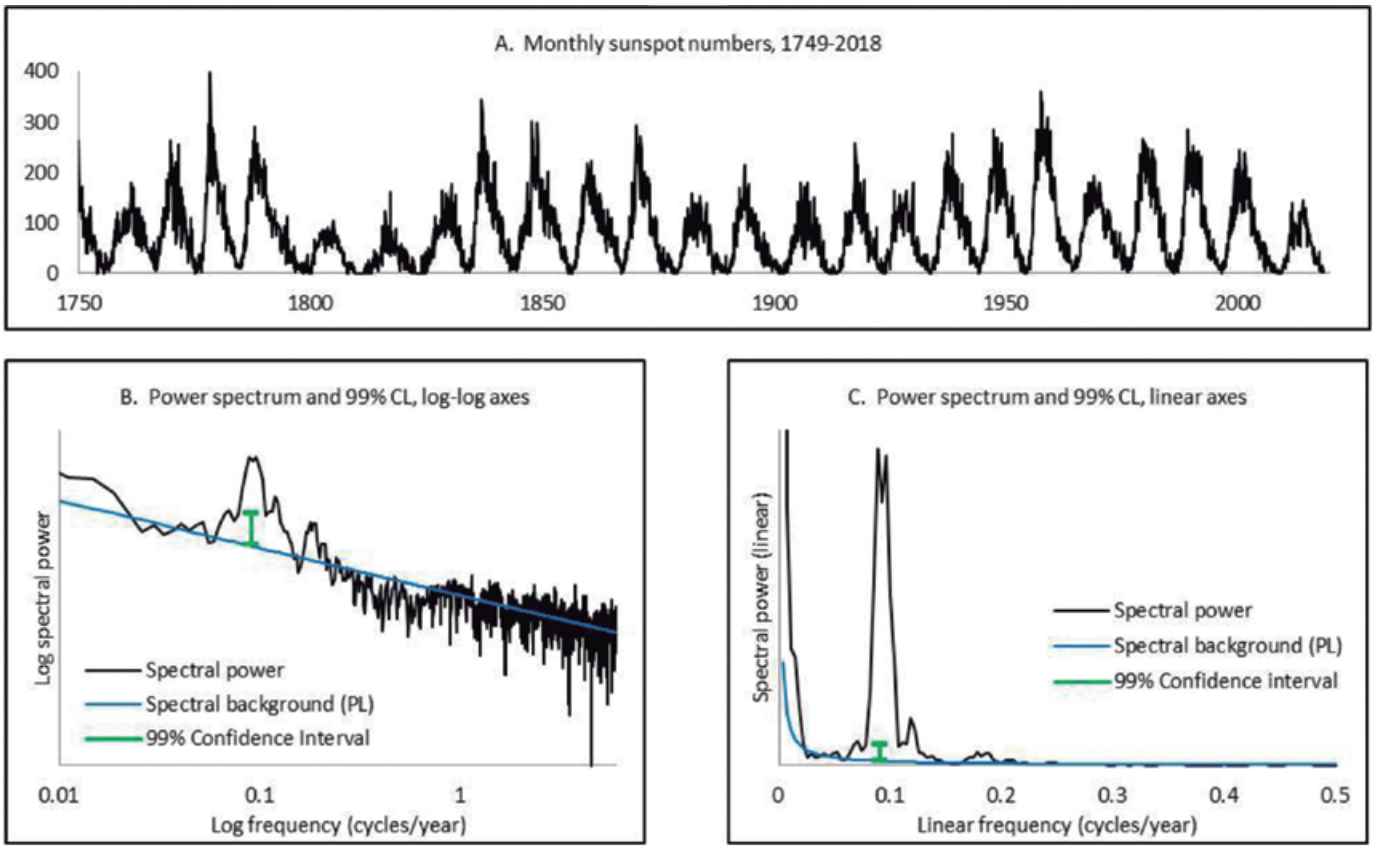

Figure 3. Single-frequency, hypothesis-based significance test: correct use of conventional method (sunspot numbers).

Figura 3. *Para una única frecuencia, test de significación basado en hipótesis: uso correcto del método convencional (número de manchas solares).*

In cyclostratigraphy, the conventional procedure generates statistical criteria (noise model and confidence threshold) for a significance test at a single frequency. Here, such a test is (correctly) applied to confirm 11-year cyclicity in sunspot numbers; the confidence interval (CI, vertical bar) is applied only at the test frequency. At this frequency, Power>>CI, and the (random) null hypothesis is rejected: the hypothesised 11-year cycle is confirmed.

A: Monthly sunspot numbers, 1749-2018 (source: WDC-SILSO, Royal Observatory of Belgium, Brussels).

B, C: Log-log and linear-axis spectrogram plots calculated using Astrochron function *mtmML96* modified to apply a power-law noise model, which is a better fit (RMS error = 14.61) to the data's spectrum than the default AR1 model (RMS=17.68). The 'robust' noise model was used to estimate a 99% confidence interval (vertical green bar) at the single frequency (0.0909 cycles/yr) corresponding to the test wavelength (11 years).

ed – affect the reliability of a statistical test. As an analogy, consider throwing several dice (Figure 2) while maintaining the expectation that a six will be seen only on one in every six throws. While this 1:6 expectation is true for any one throw of a single die, it is no longer true when the number of dice thrown is increased. The greater the number of dice, the less likely it is that a multiple throw will *not* include a six.

Similarly, in performing a statistical test, the false positive rate (e.g. 5%) may be small for a single test, but multiple repeats of the test provides multiple opportunities for a false positive result: the sixes in Figure 2 are the product of chance, and are analogous to the frequencies at which Power>CL in Figure 1.

Applying the single-die expectation to a multiple dice throw is misleading; it might, for example, lead to a conclusion that the dice are biased. Unmediated use of the conventional (ML96) power spectral calculation in cyclostratigraphy automatically supplies a single-test confidence threshold; if applied at multiple frequencies, false positive cycle identifications will be the result. This is why Figure 1 appears to indicate that this random dataset is cyclic at a number of frequencies.

For the dice, correcting the expectation can be done in two ways. Either, (1) by specifying in advance which individual die we are considering (in which case the chance of a six remains at 1:6). Or, (2) by adjusting the 1:6 single-die expectation to allow for the total number of dice cast. The options for correcting

statistical tests in cyclostratigraphy are analogous: either, (1) specify in advance at which frequency the spectral power is to be tested; or, (2) adjust the confidence threshold to allow for the number of times the significance test is to be applied.

Naïve application of uncorrected ML96 criteria at multiple frequencies can only result in multiple false positives, as in Figure 1. The ML96 single-test confidence limit can however be used, without correction, in an appropriate situation, and this is illustrated in the following section.

Conventional method: application at a single frequency

In section 3 and Figure 1, above, I showed how the customary absence of any stated objective invites misuse of the conventionally calculated confidence limit(s). Here I show how the ML96 procedure is correctly used, without any need for modification, for testing an explicit hypothesis at a single frequency. The target frequency here (0.0909 cycles per year) is the 11-year sunspot cycle.

To illustrate the principle, I first use the conventional method to test the power spectrum of the sunspot numbers themselves. The null hypothesis is that spectral power at F=0.0909 cycles/year fits a random model of the noise background at some user-defined confidence threshold; the null hypothesis applies only at that one frequency. **[Figure 3]**

Figure 3A shows historical monthly sunspot numbers from 1749 to 2018. Figure 3B/C presents the output from the conventional (cyclostratigraphic) procedure, again using the Astrochron package for simultaneous calculation of MTM power spectrum, 'robust' spectral background, and a single (95%) confidence limit. (See figure caption for further details; note that it was necessary to modify the *mtmML96* function to substitute a simple power law for the default AR1 noise model, which is a poor fit in this case. See Vaughan, Bailey and Smith (2011, Appendix B2) for the appropriate equation, and for further advice on fitting alternative noise models.)

In contrast to the multi-frequency search for significant peaks in Figure 1, a significance test is necessary at exactly one frequency in the sunspot spectrum; I have used a CL of 99% in this case. Figure 3B/C therefore shows the calculated confidence threshold only at F=0.0909, in the form of a *confidence interval*, a vertical bar connecting the noise median and the 99% confidence level at that frequency (see Weedon 2003, Figure 3.26 for another example).

At this frequency, the data's spectral power far exceeds the 99% CL (Power>>CL), and the null hypothesis (of randomness at that frequency) can be very confidently rejected; this in turn confirms the scientific hypothesis that sunspot numbers vary on an 11-year cycle.

The same test is now applied to a historical climate data-series, the (annual) Central England Temperature record (Figure 4A: details in caption). The hypothesis to be tested is the same: that there is significant cyclicity at a wavelength of exactly 11 years; the statistical test likewise starts from a null hypothesis of no cyclicity at that frequency. **[Figure 4]**

Figure 4B/C plots the results of the Astrochron calculation of a power spectrum, a 'robust' noise model, and a 95% confidence estimate for the annual CET data. (A power law model again provided a better fit than the default AR1 model – see caption.) The confidence interval (vertical green bar) is again shown at the sunspot frequency, F=0.0909 cycles/yr.

Spectral power at this frequency clearly falls short of the CL (Power<<CL): the null hypothesis is therefore accepted, with the conclusion that spectral power at this frequency is random. The (scientific) hypothesis is thus not confirmed: this dataset does not show any influence by the sunspot cycle.

However, the conventional ML96 calculation provides an apparently continuous 'confidence limit' across the entire frequency range (dashed red line in Figure 4B/C), and, as in Figure 1, this CL intersects the spectrum to define apparently significant peaks in spectral power at several (non-sunspot) frequencies. Is the CET dataset therefore cyclic at these frequencies?

Calculating and plotting a single-frequency test threshold at many frequencies is equivalent to throwing many dice: the frequencies at which Power>CL in Figure 4B/C are analogous to the sixes in Figure 2: they are the result of chance. Multiple repetition of the single-frequency test at many frequencies simply multiplies the likelihood of chance occurrence of some of the more extreme values in the noise model's probability distribution.

If the power spectrum of the CET data is to be inspected for possible cyclicity at *all* frequencies, the test threshold must be raised, to correct for the effect of statistical multiplicity; this is not optional, but is a straightforward arithmetical requirement of the statistical test. I now review the approach to such multi-frequency corrections that was introduced to cyclostratigraphy by Vaughan, Bailey and Smith (2011), and apply it both to the CET case, and to the random example of Figure 1.

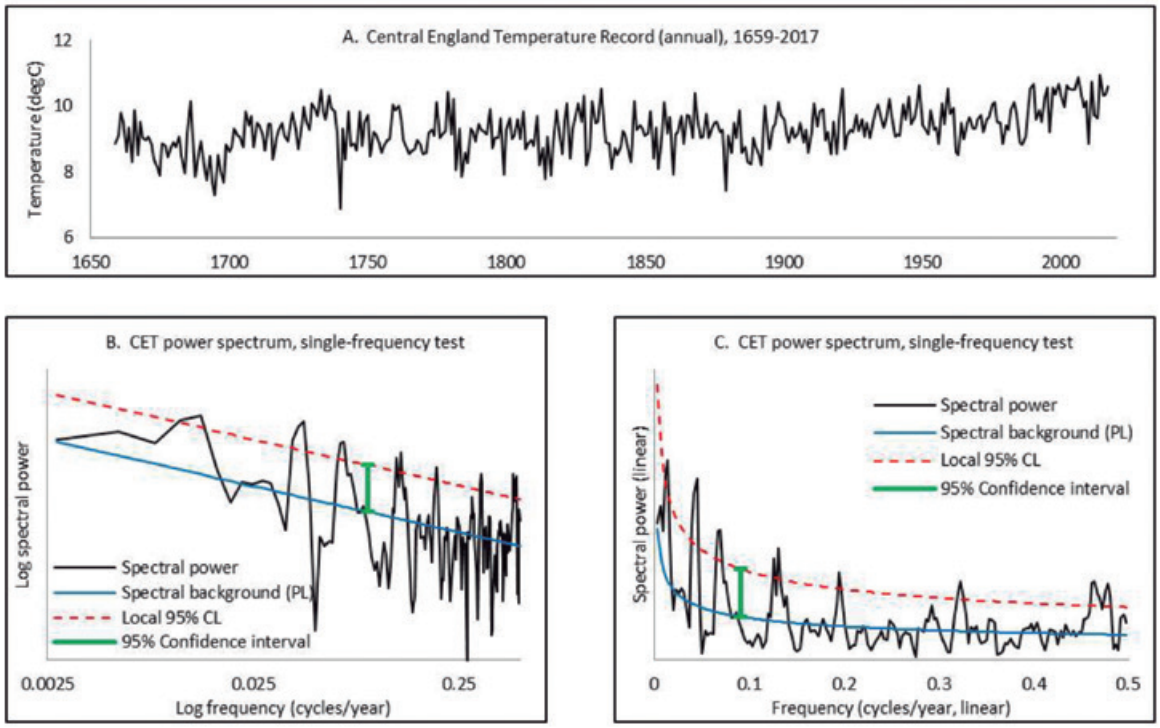

Figure 4. Incorrect extension of conventional single-frequency CL to all frequencies (CET data).

Figura 4. *Extensión incorrecta del nivel de confianza convencional para una única frecuencia a todas las frecuencias (datos CET).*

The conventional cyclostratigraphic procedure is used here to test a climate data-series for significance at the sunspot-cycle frequency (green vertical bar). At this frequency, Power<<CL, so the null hypothesis is accepted, the data are random at this frequency, and the hypothesised influence of the sunspot cycle is not confirmed. Although there is no 11-year cyclicity, the single-frequency test, conventionally plotted across all frequencies (red dashed line), finds Power>CL at several other frequencies. This is, however, misleading; Figure 5 shows that an appropriately corrected CL finds no such cycle-periods.

A: Central England Temperature (CET) record, 1659-2017, annual values (source: Met Office (UK) Hadley Centre for Climate Change).

B, C: Log-log and linear-axis spectrograms showing spectral power, 'robust' power-law spectral background and a 95% Confidence Interval (green vertical bar) at 0.0909 cycles/year, all calculated simultaneously using modified Astrochron function *mtmML96*, as for Figure 3. Dashed red line is the conventional 95% CL, which falsely suggests significance at other frequencies where Power>CL. Power law (RMS error 3.73) is a better fit to the data's spectrum than the default AR1 model (RMS=4.01).

Multiplicity: correcting for multi-frequency searches

The chance of throwing a six with a single die is 1 in 6 (16.7%); for two dice thrown together it is 11 in 36 (30.5%); the probability of seeing at least one six continues to increase rapidly with the number of dice. For the multiple throw illustrated in Figure 2, it is very unlikely that there will *not* be a six. Similarly, any uncorrected multi-frequency analysis of a random dataset (e.g. Figure 1) is very likely to include spectral peaks that exceed a single-test significance threshold by chance alone.

Posting '95% confidence' on a spectrogram suggests (qualitatively) a reassuring level of reliability; quantitatively, it predicts a false positive result (Type I error) for only one *case* in every twenty. The question of multiplicity in cyclostratigraphy turns on what is meant by 'a case'.

In Figures 3 and 4, the conventional procedure was shown to work correctly for a test of significance at a single frequency. The use of the 95% CL implies acceptance of a Type I error for one in twenty such tests; that is, one Type I error is predicted for similar analyses of twenty datasets: the *case* is the dataset. Where multiple tests are applied (or implied, as in Figure 1) within a single spectrogram, each *frequency* is a case.

The 5% false positive (FP) rate implied by a 95% CL suggests a high level of reliability; a risk of Type I errors occurring in only 1 in 20 *spectrograms* seems very acceptable. For the conventional (ML96) single-test CL, however, false positives are predicted at 1 in 20 *frequencies*. Bearing in mind the dice analogy, intuition suggests that the expected FP rate should be adjusted, such that the 1:20 FP rate applies to the whole spectrogram, instead of per frequency.

For the 1024-pt random dataset of Figure 1, a *search* for all frequencies at which Power>CL requires a null hypothesis significance test at all $N/2$=512 frequencies.

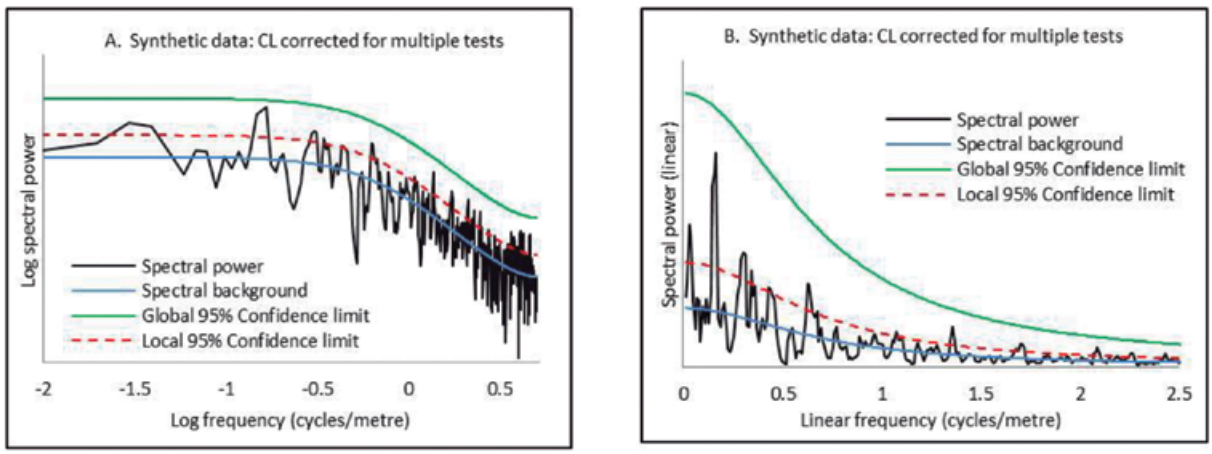

Figure 5. Correcting conventional CL for multi-frequency search (random data; compare Figure 1).

***Figura 5**. Corrección del nivel de confianza convencional para búsqueda de múltiples frecuencias (datos aleatorios, comparar con la Figura 1).*

Correction of the conventional single-test 95% confidence limit in Figure 1 allows it to be used for a multi-frequency search of the spectrogram. For the search implied in Figure 1B/C, the false positive rate must be adjusted so as to apply to the entire spectrogram. Here, using the Bonferroni correction, the 1:20 false positive rate was divided by 512, the number of frequencies in the spectrogram. The corrected 95% 'global' CL is shown in green. There is now no frequency at which PS>CL; the revised null hypothesis (of randomness at all frequencies) can be accepted, and the data are correctly confirmed as non-periodic.

A, B: As Figure 1: dashed red line is the conventional (local, i.e. single-test) 95% CL, equivalent to calculating the single-test confidence interval for all frequencies. Green line provides a 'global' 95% CL such that false positives will occur once in every 20 similar spectrograms; this is achieved by setting the 'local' (single-frequency) test threshold to 99.99%.

On average, the single-test 95% CL will be exceeded by chance at 5% of all frequencies, i.e. at 25.6 frequencies. This prediction adequately explains the number of frequencies at which Power>CL in Figure 1. Just as for the sixes that appear on the dice in Figure 2, there is no reason to invoke any explanation other than chance.

To convert the false positive (FP) rate to apply to the whole spectrogram, the Bonferroni Correction was recommended by Vaughan et al. (2011). The desired 'global' (spectrogram-wide) FP rate is divided by the number of frequencies in the spectrum to give the corrected 'local' (single-test, single frequency) FP rate. For the spectrogram in Figure 1, a 5% global FP rate is achieved by dividing 0.05 by 512, giving a corrected single-test FP rate of 0.0000977, and a confidence level of 99.99%. (While such a threshold may seem unreasonably high, it is only a simple consequence of the arithmetic of probability.)

Thus, for a simultaneous test at all 512 frequencies that will yield a false positive result for only 5% of all similar spectrograms, the (local) confidence level needed is 99.99%. This has been calculated for the Figure 1 example, as shown in Figure 5A/B. Assuming that the question was "Are there *any* frequencies at which Power>CL?", this revised CL gives the correct answer: there are no frequencies at which Power>CL. **[Figure 5]**

Figure 6A/B shows a similar correction for interrogating the CET spectrum at all frequencies. Spectral power does not intersect the corrected CL (green line) anywhere; the result is to rule out cyclicity at any frequency in this dataset (see the figure caption for more detail). **[Figure 6]**

Finally, it should be noted that correcting a single-test CL for multi-frequency application is not optional; it is a straightforward requirement of the statistical arithmetic. Arguments that such corrections are either irrelevant in cyclostratigraphy, or that they are unnecessarily 'extreme' (e.g. Hinnov, Wu and Fang 2016), stem from the conventionally casual use of confidence limits, which is exemplified in the following case study.

A case study: are confidence limits tests or guides, mandatory or optional?

I now turn to a recent cyclostratigraphic study (Ruhl, Hesselbo, Hinnov et al. 2016), in which the conventional procedure was used to analyse data-series from a rhythmically-bedded succession. CLs were calculated and displayed in this study, but contributed little to its conclusions. The subsequent Comment/Reply exchange (Smith and Bailey 2018a; Hinnov, Ruhl and

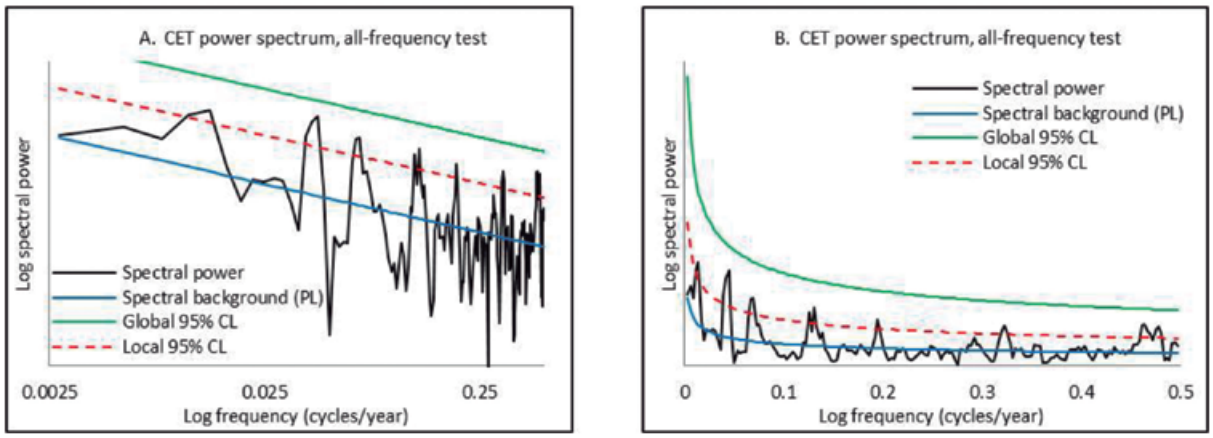

Figure 6. Correcting conventional CL for multi-frequency search (CET data; compare Figure 4).

Figura 6. *Corrección del nivel de confianza convencional para búsqueda de múltiples frecuencias (datos CET, comparar con la Figura 4).*

Correction of the conventional single-test 95% confidence limit in Figure 4 allows it to be used for a multi-frequency search of the spectrogram. For the search implied by the all-frequency extension of the single-test CL (red dashed line in Figure 4B/C), the false positive rate must be adjusted so as to apply to the entire spectrogram. Here, using the Bonferroni correction, the 1:20 false positive rate was divided by 179, the number of frequencies in the spectrogram. The corrected 95% 'global' CL is shown in green. There is now no frequency at which PS>CL; the revised null hypothesis (of randomness at all frequencies) can be accepted, with the conclusion that the data are non-periodic.

A, B: As Figure 4: dashed red line is the conventional 95% CL, the single-test confidence threshold calculated for all frequencies. The green line provides a corrected, 'global' 95% CL such that false positives will occur only once in every 20 similar spectrograms; this is achieved

Hesselbo 2018) illustrates the difference of opinion (central to the present paper) over the validity and application of confidence limits. **[Figure 7]**

I refer to the dataset analysed in Figure 4B of Ruhl, Hesselbo, Hinnov et al. (2016, Supplementary Material), but solely in order to explore the question of statistical testing as evidenced in that figure; I am not concerned here with the original authors' subsequent application of their orbital-cycle picks to 'tuning' and hence to timescale calibration. Brief details of the data and analyses can be found in the caption, and in the papers cited above.

The succession oscillates more or less rhythmically between two different lithofacies to form primary, metre-scale bedding couplets. Based on their visual observations of the borehole core, Ruhl, Hesselbo, Hinnov et al. (2016) proposed that these primary rhythms are organised into bundles of 4-5 couplets, and these into super-bundles, making a case for identifying these three orders of cyclicity with orbital precession, short eccentricity, and long eccentricity.

XRF-compositional data-series from the core 'were analysed with the 3π multi-taper method (MTM) using the Astrochron toolkit …, *with robust red noise models* (Mann and Lees, 1996) …' (Ruhl, Hesselbo, Hinnov et al. 2016, section 5.3; my italics). This description of their analytical procedure, though extremely brief, is sufficient to identify it as describing unmediated application of the conventional approach. Note that the presentation in their Figure 4B is unconventional: the usual (AR1) noise model appears (red line) in their upper panel, and the confidence limits in the middle panel, where they are used as the graph's vertical scale. Their analysis was conducted in the usual hypothesis-free (exploratory) mode.

Figure 7B/C uses a more conventional presentation to show the same 'black box' calculation of the power spectrum with default AR1 noise model, and a single confidence limit (see figure caption for more detail). Simple visual inspection reveals numerous frequencies at which Power>CL; compare Figure 1.

Ruhl, Hesselbo, Hinnov et al. (2016) were able to find spectral peaks at frequencies approximating those from their semi-quantitative visual observations of the core, but this set of frequencies has little in common with those for which Power>CL. That is, their preferred interpretation of the power spectrum largely ignores the statistical criteria. The frequencies selected by Ruhl, Hesselbo, Hinnov et al. (2016) are labelled (by wavelength) in Figure 7D/E.

In their discussion of this study, Smith and Bailey (2018a) took the statistical criteria at face value; they assumed that the noise model and confidence limits express a formal test of statistical significance. Although no objective for such a test was stated, the appearance of CLs suggests something more than a simple display of the power spectrum: an invitation to search the spectrum for significant frequencies. Multiplicity is therefore implied, and requires correction. Following Vaughan, Bailey and Smith (2011), Smith and Bailey (2018a) first applied a better-fitting noise model (to control that source of false positives), then elevated the (local) con-

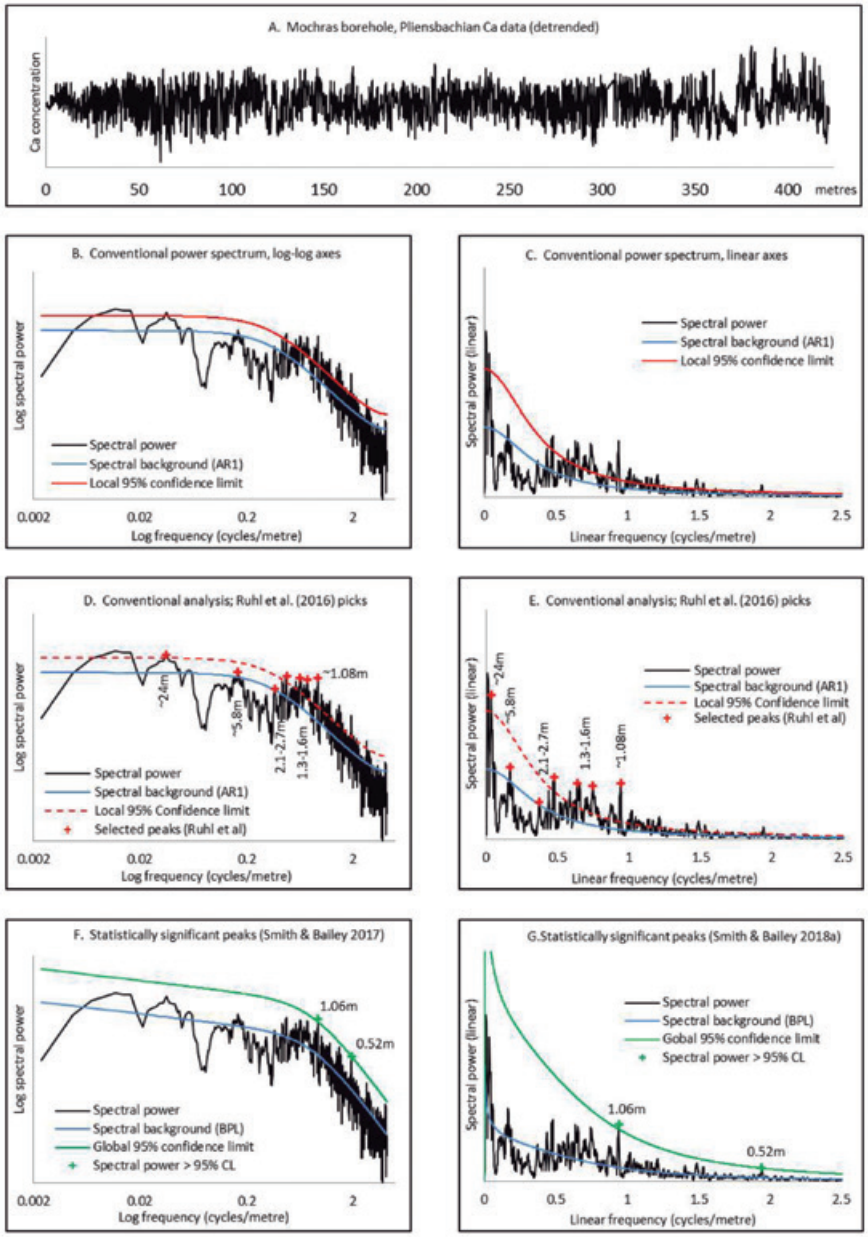

Figure 7. Conventional and corrected use of ML96-based confidence limits (Pliensbachian, Mochras, Wales).

Figura 7. *Uso convencional y uso correcto de los límites de confianza basados en ML96 (Pliensbachian, Mochras, Wales).*

This figure uses a recent case study (Ruhl, Hesselbo, Hinnov et al. 2016) to compare the statistically correct application of confidence limits with their conventional, informal use. The automatically generated statistical criteria were technically incorrect, but did not influence the selection of candidate orbital frequencies, which was based on observation of stratal patterns in the sampled core. Smith and Bailey (2018a) corrected the statistics, confirming the existence of cyclicity, but at fewer frequencies.

A: Ca concentration data through the 422 m Pliensbachian interval, Mochras borehole, N. Wales, resampled at 0.12 m intervals (3522 points) and detrended as described in Smith and Bailey (2018a).

B, C: Conventional spectral analysis simulating that of the original authors, but presented in log-log and linear-axis plots to conform with the other spectrograms in this paper (compare Ruhl, Hesselbo, Hinnov et al. 2016, Supplementary Material Figure 4B, in which the noise model appears in the top panel, and the CLs in the middle panel). My analysis simulated theirs in using the all-in-one function *mtmML96* for simultaneous calculation of power spectrum, default AR1 noise model, and default single-test CLs, of which only the 95% CL is shown here. Note that Power>CL at many frequencies (compare the random data case in Figure 1).

D, E: Same as B, C, showing frequencies (labelled by wavelength) selected by the original authors; this set of frequencies is clearly not constrained by those at which Power>CL.

F, G: Re-analysis by Smith and Bailey (2018a), who (a) found a better fitting noise model and (b) corrected the CL to allow it to be used for an all-frequency search for significant peaks with a much-reduced risk of false positives. Statistical significance (Power>CL) is attained at approximately 1 m and 0.5 m wavelengths, but not elsewhere. The noise model is a bending power law (BPL, see Appendix B, Vaughan, Bailey and Smith 2011), with index changing from 0.0 to -2.2394 at 1.0 cycles/m; the RMS error for the AR1 model is 490; for the BPL model it is 354, a substantial improvement.

fidence limit (as described in section 6 above) to correct for testing at multiple frequencies. Thus corrected, the confidence limit identifies statistically significant spectral peaks, at wavelengths of ~1 m and (marginally) ~0.5 m (Figure 7F/G, green line).

We can now compare these two different approaches to the power spectrum. Ruhl, Hesselbo, Hinnov et al. (2016) sought frequency-domain confirmation of the already-proposed hierarchy of stratification cycles, which led them to pick the local spectral peaks labelled in Figure 7D/E. This required their confidence limits to be treated as optional and non-mandatory. Smith and Bailey (2018a), taking both noise and CLs to indicate a genuine significance test, observed that both had been wrongly calculated if they were to give statistical support to a search of the spectrum. Making the appropriate corrections led to a result that was different, but which carries statistical validation.

Given that Ruhl, Hesselbo, Hinnov et al. (2016) effectively ignored their own statistical criteria, the methodological defence by Hinnov, Ruhl and Hesselbo (2018, replying to Smith and Bailey 2018a) is largely irrelevant to this particular case study; in effect, they are defending the exploratory nature of their approach, and their perceived right to treat CLs as informal guides. Their reply is nevertheless of general relevance to the wider question of how statistics are to be applied (if at all) to the results of spectral analysis in cyclostratigraphy.

Vaughan, Bailey and Smith (2011) sought to account for the typically large numbers of cycle identifications that are implied by conventionally calculated CLs. They identified two sources of false positives: incorrect noise modelling, and failure to correct for multi-frequency testing. Defence of the conventional approach centres on the claim that it is both permissible and desirable to use statistical confidence limits in a much less formal manner; this is used to justify the lax treatment of classical statistical protocols, as well as the optional, non-mandatory interpretation of the resulting CLs (Hinnov, Ruhl and Hesselbo, 2018; Hinnov, Wu and Fang, 2016; Hilgen, Hinnov, Aziz et al., 2015).

The cyclostratigraphically conventional requirement of the power spectrum in a study such as that of Ruhl, Hesselbo, Hinnov et al. (2016) is that it should yield peaks at frequencies that support a presupposition of orbital cyclicity. Given such a (scientific) hypothesis, and the desirability of confirming it quantitatively, is it not reasonable to invoke a confirmatory test of significance, in the form of a statistical confidence limit? The problem is that statistics cannot directly test a scientific hypothesis, and indiscriminate posting of incorrectly calculated CLs on cyclostratigraphic power spectra does not provide such a test. Further, it must always be accepted that one possible outcome of a properly applied statistical test is that the null hypothesis might not be rejected, implying that the proposed cyclicity cannot be statistically distinguished from random noise. Ruhl, Hesselbo, Hinnov et al. (2016) set out to confirm core-based cyclicity in the Pliensbachian of Mochras, and did so to their satisfaction. Hinnov, Ruhl and Hesselbo (2018) make it clear that no confirmatory statistical testing could be allowed to overrule this; thus, their noise modelling and confidence limits did not contribute to their conclusions.

Discussion: missing hypotheses, multiplicity, and misleading confidence

Both the random dataset (Figure 1) and the Pliensbachian case (Figure 7) exemplify the problem caused by default co-generation of a noise model and associated confidence limit(s) with every cyclostratigraphic power spectrum, and without any explicit scientific or statistical hypothesis. This practice gives the false impression that the CLs can be used as freestanding guidelines, unrestricted by the formalities of a real test of significance.

The usual absence of any explicit reference to hypothesis-testing conceals the key facts that (1) the ML96 method necessarily involves a null hypothesis, and (2) the resulting confidence test can be applied at only one frequency. Because it repeats the same (single-frequency) confidence calculation across the whole spectrum, ML96 (and its various software manifestations) appears to provide a valid significance threshold at every frequency, giving rise to the misleading situation criticised in this paper.

In cyclostratigraphy, tests of significance of spectral power are rarely possible, largely because statistical multiplicity impedes estimation of reliable confidence thresholds. It was possible to apply the conventional ML96-based method in tests for the sunspot frequency (Figure 3 and 4) (a) because the target frequency (in cycles per year) is precisely known, and (b) because the datasets tested (being historical) have unambiguous timescales (they are true time-series, unlike most stratigraphic data-series). Multiplicity arises in significance tests of cyclostratigraphic power spectra because those conditions can only rarely be met, and then only for the geologically youngest cases. In the great majority of cases, searching spectra at multiple frequencies cannot be avoided, and single-test CLs must be corrected accordingly; the alternative is to abandon statistical tests altogether.

Two additional, largely unacknowledged and entirely unresearched sources of statistical multiplicity apply to most studies in cyclostratigraphy: multiple procedural pathways, and target flexibility. Both of these are positive advantages in exploratory data analysis; but neither can be tolerated in the statistical test regime of confirmatory data analysis. Cyclostratigraphy typically operates in exploratory mode, with no need either for standardised procedures or for advance specification of target frequencies, but it is not possible to invoke statistical tests and confidence limits in such an environment of maximum flexibility.

Procedural flexibility has been described as Researcher Degrees of Freedom (Simmons, Nelson and Simonsohn 2011), and as the Garden of Forking Paths by Gelman and Loken (2014): potential analytical pathways, and hence opportunities for chance positive results, proliferate at every decision-point in the procedure. Decision points in cyclostratigraphic investigations include: section selection and sampling strategy; data pre-processing (outlier removal, interpolation, detrending); choice of spectral method, software package, and parameter settings; noise estimation and confidence levels; plotting parameters (log versus linear axes, full spectrum or partial frequency scale); the list is long. In the exploratory mode of the conventional approach it seems entirely reasonable to adapt the procedure to suit each individual dataset (see the Methods section of any cyclostratigraphic study). This, however, implies many *potential* procedural pathways, each of which adds to the range of possible outcomes and therefore degrades confidence thresholds if confirmatory tests of significance are to be applied. A study by Carp (2012) demonstrated the complexity of trying to adjust statistical tests for all such sources of flexibility in data collection and analysis, though this was not in cyclostratigraphy where no such study has yet been undertaken. Procedural flexibility, incompatible with statistical testing, is essential to the exploratory phase of data analysis, reinforcing my view that cyclostratigraphic spectral analysis operates very largely in exploratory mode.

Target flexibility is also desirable in exploratory analysis, and is also essential to conventional procedures in cyclostratigraphy; it is likewise difficult to reconcile with rigorous tests of significance. Conventionally, the power spectrum is inspected for candidate frequencies, either in an entirely open-ended way (as simulated in Figure 1), or with the aim of matching visual observations to a plausible suite of orbital frequencies (as in the Pliensbachian case, Figure 7). This approach is necessary because target frequencies cannot be pre-specified (especially in the depth domain) because of time-scale uncertainties; this is particularly true if time-depth calibration is the objective of the study and circularity is to be avoided (Bailey 2009). Such flexibility, however, is incompatible with the confirmatory analysis that is implied by plotting confidence limits with the power spectrum, because the statistics cannot cope with the inevitably high level of multiplicity. Retrospective target specification is a particularly rich source of statistical multiplicity, carrying the risk of 'data-mining', 'p-hacking', or 'torturing the data until it confesses' (Nuzzo 2015). In the approach to cyclostratigraphic analysis used in the Average Spectral Misfit (Meyers and Sageman, 2007) and TimeOpt (Meyers 2019b) methods, the user is actively encouraged to experiment across a wide range of possible solutions until satisfied with the outcome (Graham Weedon, pers. comm.); such deliberate appeal to multiplicity must cast suspicion on such methods, which should be comprehensively tested with random datasets before being relied upon to analyse any real data.

The overall effect of all sources of multiplicity is the tendency to vastly expand the range of possible outcomes, making reliable calibration of confidence thresholds impossible. The implication is paradoxical: treating CLs as flexible destroys any value they may originally have had; by treating them 'as a guide', they effectively cease to exist. The conventional all-frequency calculation of CLs appears convenient, because it provides a test at any frequency. Yet it is highly misleading because, once a single-frequency test has been conducted, the CL no longer applies at any other frequency in the same spectrogram without correction; the CET case (Figures 4 and 6) is an example.

For completeness, I mention one further modifier of confidence thresholds. Prior probability refers to the *a priori* likelihood of occurrence of any effect under investigation. In cyclostratigraphy, there is a clear need to consider (in advance of analysis) the underlying probability that the depositional system in question is capable of recording climate change at orbital periodicities (Bailey 2009). As with medical diagnosis, the burden of proof is necessarily greater where the occurrence of the target effect is *a priori* less likely (Ioannidis 2005, Nuzzo 2014). There is surely potential for erecting a semi-quantitative scale (or at the very least, an ordinal scale) of depositional environments, from those least likely to preserve such a record to the most likely; such a scale might be difficult to apply, but would at least draw attention to this issue. Bayesian statistics is available to deal with the conditional probabilities that arise from variable priors of this kind.

Conclusions and recommendations

My title suggests problems with the confidence limits that routinely appear on cyclostratigraphic power spectra: Why are they so often calculated and plotted, yet so rarely adhered to? Why have they been so strongly defended against technically correct criticism? Is it possible to clarify and resolve these differences? In this paper, I have identified technical problems with (1) the conventional, standardised nature of power spectral analysis in cyclostratigraphy; (2) confusion over the role of hypotheses; and (3) the unavoidable impact of statistical multiplicity. These three are closely related to each other; reliance on conventionalised procedures has obscured the unavoidable role of hypotheses in significance-testing, encouraging the spurious application of confirmatory statistics in the 'forking paths' (high multiplicity) environment of exploratory-mode data analysis.

Resolution of the technical issues will be enabled by (1) avoiding the 'black box' calculations (mainly based on ML96) that automatically embellish every power spectrum with inappropriate noise models and confidence limits; (2) appreciation of the roles of scientific and statistical (null) hypotheses, and hence of the distinction between the exploratory (hypothesis-forming) and confirmatory (hypothesis-testing) stages of data analysis; and (3) awareness of actual and potential sources of statistical multiplicity and their effect on confidence thresholds.

The function of most cyclostratigraphic power spectra is necessarily exploratory; statistical tests are therefore not appropriate, and reliable confidence limits cannot be calculated. CLs should not be plotted unless they can be correctly calculated for a specific and explicit null hypothesis significance test; there is no such thing as a context-free confidence limit. 'Black box' software packages should carry appropriate warnings, and any cycle identifications that depend on conventionally calculated CLs should be challenged.

In cases for which confirmatory statistical tests can be justified, based on a properly formulated null hypothesis, the risk of false positive cycle identifications (Type I errors) must be acknowledged and managed, not dismissed and ignored. False positive rates can generally be predicted; they should be confirmed using synthetic (random) datasets. Statistical tests must be allowed to 'fail'; a null hypothesis that cannot be rejected must be accepted as evidence that a local peak is not statistically distinguishable from random. As in the Pliensbachian example above, this need not invalidate a conclusion of orbital cyclicity at that wavelength, but it does exclude statistical support.

Vaughan, Bailey and Smith (2011) showed that poorly fitting noise models can raise false positive rates. Their observations have now been accepted, and Weedon (this volume) has recently proposed a new and entirely empirical approach. Further work is needed on this and other methods, in order to establish model-fitting protocols for general application in cyclostratigraphy.

Further research is also required into appropriate corrections for statistical multiplicity in cyclostratigraphy. I have kept to the Bonferroni correction in the examples herein (Figures 5 and 6), because it is intuitive, easy to apply, and conservative (which is appropriate for minimisation of Type I errors in random data, for example). Criticisms of this particular correction (e.g. Hilgen, Hinnov, Aziz et al. 2015; Hinnov, Wu and Fang 2016), are fair but are not a reason for denial of the underlying need for multiplicity corrections. Crampton, Meyers, Cooper et al. (2018), and Weedon, Page and Jenkyns (2019) have now started to use and recommend other methods, particularly that proposed by Benjamini and Hochberg (1995), which is now very widely used in many other sciences.

Standardisation of analytical procedure (from sampling to target specification) is highly desirable if statistical testing is to be reliable, especially for comparison between datasets; Weedon (this volume) has set a clear example in this regard. Protocols for a preferred approach to exploratory analysis, without any reference to statistics but with hypothesis-formulation as the goal, would also be helpful.

If statistical significance is to be the basis of reliable cycle-period identification, the procedure for the confirmatory (statistical) stage of analysis should be specified in advance, and should ideally be conducted on a new dataset, independent of that used for the exploratory stage; these are widely regarded as minimum standards in most applications of statistics (Munafò, Nosek, Bishop et al. 2017). The practice in cyclostratigraphy of relying on a single dataset both for proposing an orbital hypothesis and for invoking statistics to confirm it, risks circular reasoning (Bailey 2009). It is unfortunate that cancellation of the second Mochras borehole has deprived cyclostratigraphy of a rare opportunity to conduct rigorous confirmatory statistics on the essentially exploratory results from Mochras-1 (Ruhl, Hesselbo and Hinnov, 2016). Cyclostratigraphers are unlikely to welcome the suggestion that they should collect and analyse every succession twice over, but such are the requirements of proper statistical practice.

There is considerable scope for the introduction of Bayesian statistics in cyclostratigraphy, where the method proposed by Vaughan (2010) has received little

attention. Weedon (this volume) makes a strong case for their use, and my plea for routine consideration of prior probabilities also requires Bayesian methods.

Reproducibility is currently a major issue in many sciences (Munafò, Nosek, Bishop et al. 2017; National Academies of Sciences, Engineering, and Medicine, 2019), and cyclostratigraphy should not be an exception to the absolute requirement that all data and methods should be completely open, and freely available.

This contribution to cyclostratigraphy has necessarily been critical, and could be interpreted as essentially negative. However, astrochronology (and the geological time-scale) need and deserve a supply of dependable identifications of orbital cycle-periods in stratigraphic successions; the currently relaxed attitude to statistical testing in cyclostratigraphy does not achieve the required level of reliability. Thirty years since my first publication in cyclostratigraphy, this is likely to be my last; if I have succeeded in clarifying (for some, at least) the role, the requirements, the operation, and the limitations of significance tests in power spectral analysis, then this contribution will – I hope – have been a positive one.

Acknowledgements

Walther Schwarzacher, to whose memory this volume is dedicated, was an invaluable source of information, advice and wisdom in my early days of working in cyclostratigraphy, in BP Research in the 1980s. He contributed papers at symposia that I organised, and fed me a memorable Wiener Schnitzel at his home in Belfast. He was generous with his thoughts, and a true gentlemen in the very highest central European tradition. Queen's Belfast (the university that responded to BP's gift to Walther of a desktop PC by asking who was to pay for the electricity) was more fortunate than it knew to have hosted him for so many decades; he will be sorely missed.

I am profoundly grateful to Graham Weedon for taking so much trouble over a previous version of this manuscript that I was left with little choice but to do a complete re-write; I believe that the result is a substantial improvement. Robin Bailey, a co-author of several relevant papers, has been an invaluable touchstone over the past two decades, and his comments have also helped to improve this paper. Simon Vaughan took time out from astrophysics and university duties a few years ago to cast his very critical eye over the way statistics had developed in cyclostratigraphy ("They can't do that!"), and, despite our detractors, the subject has benefited from his vastly superior knowledge and understanding of statistical methods, especially of their constraints and limitations. Cedric Griffiths and John Tipper have both, over many years, endured my sermons on statistics in cyclostratigraphy, among entertaining and informative discussions of many topics in mathematical stratigraphy.

Linda Hinnov, a formidable foe with far more experience of real datasets than I will ever have, has contributed materially to this investigation through personal correspondence as well as through published Replies to our critical Comments; thank you. Stephen Hesselbo kindly took time out from planning the Mochras-2 borehole to answer my challenge that most confidence limits in cyclostratigraphy are misplaced; it was his remark "But it must mean something!" that (although he was wrong) set me on the path that led to this paper; thank you, too.

References

Andrews, S.D., Cornwell, D.G., Trewin, N.H., Hartley, A.J. and Archer, S.G., 2018. Reply to the discussion on 'A 2.3 Million Year Lacustrine Record of Orbital Forcing from the Devonian of Northern Scotland' (*Journal of the Geological Society,* 173, 474–488). *Journal of the Geological Society*, 175, 563, 1 December 2017. London.

Bailey, R. J. 2009. Cyclostratigraphic reasoning and orbital time calibration. *Terra Nova*, 21(5), 340–351.

Benjamini, Y. and Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289-300.

Carp, J., 2012. The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage*, 63(1), 289-300.

Crampton, J.S., Meyers, S.R., Cooper, R.A., Sadler, P.M., Foote, M. and Harte, D., 2018. Pacing of Paleozoic macroevolutionary rates by Milankovitch grand cycles. *Proceedings of the National Academy of Sciences*, 115(22), 5686-5691.

Da Silva, A.C., Dekkers, M.J., De Vleeschouwer, D., Hladil, J., Chadimova, L., Slavík, L. and Hilgen, F.J. 2019. Millennial-scale climate changes manifest Milankovitch combination tones and Hallstatt solar cycles in the Devonian greenhouse world (Forum Reply). *Geology*, 47(10), 489-490.

Gelman, A. and Loken, E., 2014. The statistical crisis in science: data-dependent analysis – a "garden of forking paths" – explains why many statistically significant comparisons don't hold up. *American Scientist*, 102(6), 460-466.

Gong, Z. and Kodama, K.P., 2018. Reply to the comment on``Rock magnetic cyclostratigraphy of the Doushantuo

Formation, South China and its implications for the duration of the Shuram carbon isotope excursion''. (Gong, Z., Kodama, K.P. and Li, Y.X. *Precambrian Research*, 289 (2017), 62-74). *Precambrian Research*, 310, 467-470.

Hilgen, F.J., Hinnov, L.A., Aziz, H.A., Abels, H.A., Batenburg, S., Bosmans, J.H., de Boer, B., Hüsing, S.K., Kuiper, K.F., Lourens, L.J. and Rivera, T., 2015. Stratigraphic continuity and fragmentary sedimentation: the success of cyclostratigraphy as part of integrated stratigraphy. *Geological Society Special Publications*, 404(1), 157-197. London.

Hinnov, L.A., Ruhl, M.R. and Hesselbo, S.P., 2018. Reply to the Comment on "Astronomical constraints on the duration of the Early Jurassic Pliensbachian Stage and global climatic fluctuations" (Ruhl et al., *Earth and Planetary Science Letters*, 455 (2016) 149–165). *Earth and Planetary Science Letters,* 481, 415-419.

Hinnov, L.A., Wu, H. and Fang, Q., 2016. Reply to the comment on "Geologic evidence for chaotic behavior of the planets and its constraints on the third-order eustatic sequences at the end of the Late Paleozoic Ice Age" by Qiang Fang, Huaichun Wu, Linda A. Hinnov, Xiuchun Jing, Xunlian Wang, and Qingchun Jiang (*Palaeogeography Palaeoclimatology Palaeoecology* 400 (2015) 848–859). *Palaeogeography, Palaeoclimatology, Palaeoecology*, 461, 475-480.

Howe, T.S., Corcoran, P.L., Longstaffe, F.J., Webb, E.A. and Pratt, R.G., 2018. Response to the discussion on``Climatic cycles recorded in glacially influenced rhythmites of the Gowganda Formation, Huronian Supergroup'' (Precambrian Research, 286 (2016), 269-280). *Precambrian Research*, 316, 327.

Ioannidis, J.P., 2005. Why most published research findings are false. *PLos Medicine*, 2(8), e124.

Mann, M.E. and Lees, J.M., 1996. Robust estimation of background noise and signal detection in climatic time series. *Climatic Change*, 33(3), 409-445.

Meyers, S.R., 2019a. Astrochron: An R package for astrochronology (version 0.9). *https://cran.r-project.org/web/packages/astrochron/index.html*

Meyers, S.R., 2019b. Cyclostratigraphy and the problem of astrochronologic testing. *Earth-Science Reviews*, *190*, 190-223.

Meyers, S.R. and Sageman, B.B., 2007. Quantification of deep-time orbital forcing by average spectral misfit. *American Journal of Science,* 307(5), 773-792.

Munafò, M.R., Nosek, B.A., Bishop, D.V., Button, K.S., Chambers, C.D., Du Sert, N.P., Simonsohn, U., Wagenmakers, E.J., Ware, J.J. and Ioannidis, J.P., 2017. A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 1-9.

National Academies of Sciences, Engineering, and Medicine, 2019. *Reproducibility and Replicability in Science*. National Academies Press, Washington DC.

Nuzzo, R., 2014. Scientific method: statistical errors. *Nature News*, 506 (7487), 150.

Nuzzo, R., 2015. How scientists fool themselves – and how they can stop. *Nature*, 526 (7572), 182-185.

Ruhl, M., Hesselbo, S.P., Hinnov, L., Jenkyns, H.C., Xu, W., Riding, J.B., Storm, M., Minisini, D., Ullmann, C.V. and Leng, M.J., 2016. Astronomical constraints on the duration of the Early Jurassic Pliensbachian Stage and global climatic fluctuations. *Earth and Planetary Science Letters*, 455, 149-165.

Schwarzacher, W., 1975. *Sedimentation Models and Quantitative Stratigraphy*. Developments in Sedimentology, 19. Elsevier, 382 pp.

Simmons, J.P., Nelson, L.D. and Simonsohn, U., 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359-1366.

Smith, D.G., 2019. Millennial-scale climate changes manifest Milankovitch combination tones and Hallstatt solar cycles in the Devonian greenhouse world (Forum Comment: Da Silva, A.C., Dekkers, M.J., De Vleeschouwer, D., et al., *Geology*, 47 (2019), 489-490). *Geology*, 47(10), 488.

Smith, D.G. and Bailey, R.J., 2017a. Discussion on 'A 2.3 million year lacustrine record of orbital forcing from the Devonian of northern Scotland' (*Journal of the Geological Society*, 173, 474–488). *Journal of the Geological Society*, 175, 561-562. London.

Smith, D.G. and Bailey, R.J., 2017b. Discussion on 'Orbital calibration of the late Campanian carbon isotope event in the North Sea' (*Journal of the Geological Society,* 173, 504–517). *Journal of the Geological Society*, 175, 564-565. London.

Smith, D.G. and Bailey, R.J., 2018a. Comment on "Astronomical constraints on the duration of the Early Jurassic Pliensbachian Stage and global climatic fluctuations" (*Earth and Planetary Science Letters* 455 (2016), 149-165). *Earth and Planetary Science Letters*, 481, 412-414.

Smith, D.G. and Bailey, R.J., 2018b. Discussion: Howe, T.S., Corcoran, P.L., Longstaffe, F.J., Webb, E.A., & Pratt, R.G. (2016). Climatic cycles recorded in glacially influenced rhythmites of the Gowganda Formation, Huronian Supergroup (*Precambrian Research*, 286, (2016), 269-280). *Precambrian Research*, *316*, 324-326.

Smith, D.G. and Bailey, R.J., 2018c. Comment on "Rock magnetic cyclostratigraphy of the Doushantuo Formation, South China and its implications for the duration of the Shuram carbon isotope excursion". (Gong, Z., Kodama, K.P. and Li, Y.X., *Precambrian Research*, 289 (2017), 62-74). *Precambrian Research* 310, 463-466.

Smith, D.G., Bailey, R.J. and Vaughan, S., 2016. Comment on "Geologic evidence for chaotic behavior of the planets and its constraints on the third-order eustatic sequences at the end of the Late Paleozoic Ice Age" by

Fang, Q., Wu, H., Hinnov, LA, Jing, X., Wang, X., and Jiang, Q. (*Palaeogeography, Palaeoclimatology, Palaeoecology*, 440 (2015), 848–859). *Palaeogeography, Palaeoclimatology, Palaeoecology*, 461, 472-474.

Thibault, N. and Perdiou, A. 2018. Reply to the discussion on 'Orbital calibration of the late Campanian carbon isotope event in the North Sea' (*Journal of the Geological Society*, 173 (2016), 504–517). *Journal of the Geological Society*, 175, 566-567, 2018. London.

Tukey, John W. 1977, *Exploratory Data Analysis*, Reading, Mass. Addison-Wesley.

Vaughan, S., 2013. *Scientific Inference: Learning from Data*. Cambridge University Press, 224 pp.

Vaughan, S., 2010. A Bayesian test for periodic signals in red noise. *Monthly Notices of the Royal Astronomical Society, 402*(1), 307-320.

Vaughan, S., 2013. *Scientific Inference: Learning from Data*. Cambridge University Press, 224 pp.

Vaughan, S., Bailey, R.J. and Smith, D.G. 2011. Detecting cycles in stratigraphic data: spectral analysis in the presence of red noise. *Paleoceanography* 26, doi: 10.1029/2011PA002195.

Weedon, G.P., 2003. *Time-Series Analysis and Cyclostratigraphy: Examining Stratigraphic Records of Environmental Cycles*. Cambridge University Press, 259 pp.

Weedon, G.P., Page, K.N. and Jenkyns, H.C., 2019. Cyclostratigraphy, stratigraphic gaps and the duration of the Hettangian Stage (Jurassic): insights from the Blue Lias Formation of southern Britain. *Geological Magazine*, 156 (9), 1469-1509.

Weedon, G.P., 2020. Confirmed detection of Palaeogene and Jurassic orbitally-forced sedimentary cycles in the depth domain using False Discovery Rates and Bayesian probability spectra. *Boletin Geológico y Minero, 131, 2, 207-230.*

Recibido: julio 2019
Revisado: noviembre 2019
Aceptado: enero 2020
Publicado: marzo 2021