P-hacking, HARKing and confirmation bias in cyclostratigraphic spectral analysis

David G. Smith

(Independent consulting geologist)

Address: 15 Stratton Terrace, Truro TR1 3EW, U.K.

email: d.g.smith@talktalk.net

Preprint of a paper submitted for peer review to Matters Arising, *Nature Communications,* 5 July 2022. Exact content may be different in subsequent versions of this manuscript. This preprint was submitted, with the addition of a short Abstract and without peer review, to ESSOAr, 15 July 2022. Feedback and discussion will be welcomed; please contact the author.

Supplementary material: the data used in Figure 1 are available in the following supplementary file:

SMITHdg_MattersArising(ZHAOea)_SUPPL-DATA.xlsx.

Submitted to *Nature Communications,* MATTERS ARISING, 5 July 2022

P-hacking, HARKing and confirmation bias in cyclostratigraphic spectral analysis

David G. Smith[1]

[1]15 Stratton Terrace, Truro TR1 3EW, U.K., d.g.smith@talktalk.net

ARISING FROM

Zhao et al. *Nature Communications* https://doi.org/10.1038/s41467-022-29651-4 (2022)

ABSTRACT

A simple statistical test is used in cyclostratigraphy to discover candidate orbital frequencies in power spectra of climate proxy data-series. In published studies at least, this test never fails to find multiple frequencies, at high levels of statistical significance (e.g. p<0.01). However, the same method finds similarly high statistical significance at similar numbers of frequencies in random, simulated datasets. The problem lies with the standardised application of the test, which is linked to MTM spectral analysis in a one-step procedure that is readily accessible through specialist software packages. This procedure presents confidence limits as if they were context-free, but statistical tests are necessarily tied to specific (null) hypotheses. The test as used in cyclostratigraphy is calibrated for application at a single frequency, but it is routinely used as if applicable at all frequencies, a practice that invokes the statistical multiple comparisons problem and which largely explains the inadvertent conversion of noise to signal when applied to random datasets. This general problem is addressed here with reference to a specific recently published case.

The results reported by Zhao et al.[1] depended on a procedure that finds statistically significant peaks in power spectra of random data – their method finds regular cycles where none exist. Figures 1a and 1b present applications of the same procedure to, respectively, their data-series, and a random simulation of it; both plots show power exceeding the 99% confidence limit at numerous frequencies. In Fig. 1a, "The … power spectrum … revealed numerous peaks … above the 99% confidence level" (Zhao et al.[1] Supplementary Note 3, referring to Supplementary Fig. 4a). Fig. 1b should have no significant peaks, because the data are random. The false significance arises from statistical multiplicity (the multiple comparisons problem, or *p*-hacking). Because this also applies to Fig. 1a, most or all of Zhao et al.'s frequency picks must be incorrect. While there are implications for their study's particular conclusions, the following comments should be taken to apply more generally, to the widespread use of invalid statistical tests in cyclostratigraphic spectral analysis.

Zhao et al.'s data-series comprised XRF measurements of Al concentration at 1 mm spacing through ~74 m of the Cambrian (~500 Ma) Alum Shale Formation in a cored borehole from southern Sweden. Their objective was high-resolution calibration of part of Cambrian time. They used spectral analysis to seek patterns of cyclicity (in stratigraphic depth) that might indicate orbitally-forced (Milankovitch) cycles of paleo-environmental change, expected as peaks in spectral power against a background of aperiodic red noise. Their procedure followed convention: MTM spectral analysis accompanied by estimation of spectral background, from which confidence limits (CLs) were calculated and plotted with the power spectrum, as in Fig. 1a, b. A CL expresses a probabilistic test of local power against a null hypothesis of randomness, and is computed from the noise model using chi-square sampling theory. Confidence limits assess the probability that some observed value is a chance result. For the one-sided probability distribution relevant to a power spectrum, a 99% CL (for example) indicates the minimum observed value for which the probability of chance occurrence is less than 1% ($p<0.01$). (Note that 99% is *not* the probability of the underlying hypothesis being

51 true[2,3].) Fig. 1b demonstrates that the CLs, as calculated through the standard procedure, do not do
52 this correctly.
53
54 As in nearly all cyclostratigraphic investigations, it was not possible to erect a statistically testable
55 (null) hypothesis. This is because the target frequencies are so uncertain, and because of the
56 computationally difficult nature of stratigraphic data, in which stratigraphic depth is a poorly
57 constrained proxy for time. Accordingly, and consistent with the absence of visible stratification
58 cycles in their succession, Zhao et al. made no specific predictions of cyclic frequencies; their analysis
59 was, rather, a spectrum-wide *search* for frequencies of possible interest. Their analysis was
60 therefore not a null hypothesis significance test (NHST), yet their identifications of periodic
61 frequencies explicitly depended on the 99% CL, which can only represent a NHST. Their analysis was
62 thus a self-contradictory combination of hypothesis-dependent significance criteria with a
63 hypothesis-free search[4]: this is the central paradox of conventional cyclostratigraphic practice.
64
65 This standardised approach misuses a method (ML96[5]) that was developed to discriminate
66 (anthropogenic) trends from both cyclical patterns and random information in recent climate data[6];
67 it calculates confidence limits by default. The method's authors acknowledged that these CLs lead to
68 false positive cycle detections, but this was unimportant in their quest for the trend. Fig. 1b
69 demonstrates, however, that false positives are a major problem when the method is instead used
70 to search for evidence of cyclicity. As conventionally calculated, the ML96 CLs are correct for
71 application at one frequency only; false positive detections are a simple arithmetical consequence of
72 application at multiple frequencies, and >3,000 frequencies are represented in Figs 1a and b.
73 Corrections for multiple application are available[7,8] but are practised only exceptionally in
74 cyclostratigraphy[9,10]. Fig. 1b shows (dashed line) a 99% CL properly calculated for simultaneous use
75 at all frequencies; it correctly identifies no cyclic frequencies in this random dataset.
76

Statistical multiplicity is more deceptive, and less easily quantified[11], when it arises from procedural flexibility ('Researcher Degrees of Freedom'[12]) and from post-analytical target-setting, both of which apply here. Standard protocols remain non-existent in cyclostratigraphy; conventional practice assumes that all procedures may be adjusted without compromising significance. Zhao et al. accordingly adapted their pre-processing (smoothing and detrending) parameters to their data, and made their own (undeclared) choices of spectral analysis settings. Target-setting was based on clusters of 'significant' peaks, which were used to define broad frequency intervals (with the only constraint that their mutual ratios resemble those of the four main orbital periods – the coloured bands in Zhao et al. Sup. Fig. 4a); 'significant' frequency peaks falling outside these bands were rejected from further consideration. Such post-analytical target-setting has been called HARKing, or Hypothesising After Results Known[13]. The effect of such flexibility in both the conduct and interpretation of the statistical tests is to increase the range of *potential* scenarios (the 'Garden of Forking Paths'[14]), and hence the level of multiplicity. Together, procedural flexibility, *p*-hacking and HARKing eliminate any validity from the conventional CLs, inviting confirmation bias.

Statistical tests are valid and reliable only under proper procedures. Zhao et al.'s study exemplifies the widespread appearance in cyclostratigraphy of statistical thresholds generated without regard to any null hypothesis[4]. The illusion that significance can be quantified without reference to a specific hypothesis is reinforced by the automated calculation of confidence limits through implementations of ML96 in specialist software packages. Further, published precedents include no cases in which the conventionally calculated CLs prove absence of cyclicity. Instead, a majority of cases (including Zhao et al.) have reported recovery of the full suite of long and short eccentricity, obliquity and precession cycle periods, regardless of the inherent improbability of such a result. Prior probability is essential for statistical validity, even in stratigraphy where it is impossible to estimate quantitatively. Qualitative considerations confirm the minimal likelihood of achieving such a faithful orbital recording: after translation and corruption of the orbital signal through multiple climatic,

sedimentary, and post-depositional processes, reliable recording of numerically coherent information in the resulting strata can only be exceptional. Bad statistics is a much less implausible explanation for the numerous cycle detections implied by Fig. 1 (and in hundreds more published cases).

Zhao et al.'s Supplementary Fig. 4a was the *origin* of their cyclic frequencies. Before the adoption of ML96, spectral analysis was optional in cyclostratigraphy (and statistics more so). Classical observation and measurement, with sedimentological and geochemical analyses, were the primary means of investigating observed stratification cycles[15]. Dependence on the ML96 plot is now the default; its reliably positive results seem to obviate any concerns about hypotheses, prior probabilities, or *p*-hacking. Zhao et al.'s analysis rewarded their expectations of multi-frequency periodicity, in turn justifying cherry-picking from these to populate a retro-fitted multi-frequency target. In fact, the CLs are deprived of any possible meaning because: cyclicity (though not impossible) is unlikely *a priori*; the null hypothesis built into the conventional analysis applies at only one frequency; and flexible procedures invoke unquantifiable multiplicity from additional sources. The conventional view, implicit in Zhao et al. (and explicit elsewhere[16]), is that the uncertainties inherent in the data justify a casual approach to the interpretation of confidence limits. I argue instead that the uncertainties in both the data and the objective require more statistical rigour, not less.

Data availability

All data used to construct Figures 1a and 1b are included in the Supplementary Information (see below).

REFERENCES

1. Zhao, Z. et al. Synchronizing rock clocks in the late Cambrian. *Nat. Commun.* **13**, 1-11 (2022).
2. Nuzzo, R. Statistical errors. *Nature* **506**, 150-152 (2014).
3. Anon. Editorial: The correct use of statistics is not just good for science — it is essential. *Nature* **506**, 131-2 (2014).
4. Smith, D.G. Misplaced confidence: limits to statistical inference in cyclostratigraphy. *Bol. Geol. Min.* **131**, 291-307 (2020).
5. Mann, M.E. & Lees, J.M. Robust estimation of background noise and signal detection in climatic time series. *Climat. Change*, **33**, 409-445 (1996).
6. Mann, M.E., Bradley, R.S. & Hughes, M.K. Global-scale temperature patterns and climate forcing over the past six centuries. *Nature*, **392**, 779-787 (1998).
7. Vaughan, S., Bailey, R.J. & Smith, D.G. Detecting cycles in stratigraphic data: spectral analysis in the presence of red noise. *Paleoceanography* **26**, PA4211, doi:10.1029/2011PA002195 (2011).
8. Streiner, D.L. Best (but oft-forgotten) practices: the multiple problems of multiplicity — whether and how to correct for many statistical tests. *Amer. J. Clinic. Nutrition* **102**, 721-728 (2015).
9. Crampton, J.S. et al. Pacing of Paleozoic macroevolutionary rates by Milankovitch grand cycles. *Proc. Nat. Acad. Sci.* **115**, 5686-5691 (2018).
10. Weedon, G.P., Page, K.N. & Jenkyns, H.C. Cyclostratigraphy, stratigraphic gaps and the duration of the Hettangian Stage (Jurassic): insights from the Blue Lias Formation of southern Britain. *Geol. Mag.* **156**, 1469-1509 (2019).
11. Carp, J. The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage* **63**, 289-300 (2012).

12. Simmons, J.P., Nelson, L.D. & Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359-1366 (2011).
13. Nuzzo, R. Fooling ourselves. *Nature* **526**, 182-185 (2015).
14. Gelman, A. & Loken, E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Dept Statistics, Columbia University* **348,** 1-17 (2013). http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
15. Schwarzacher, W. C*yclostratigraphy and the Milankovitch Theory*. Elsevier, Amsterdam (1993).
16. Hinnov, L.A., Wu, H. & Fang, Q.. 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". *Palaeogeogr., Palaeoclim., Palaeoecol.*, **461**, 475-480 (2016).

Competing interests

The author declares no competing interests.

Additional information

The data displayed in Figure 1, and the methods used to generate these plots, can be found in the Excel spreadsheet available with this preprint.

FIGURE 1

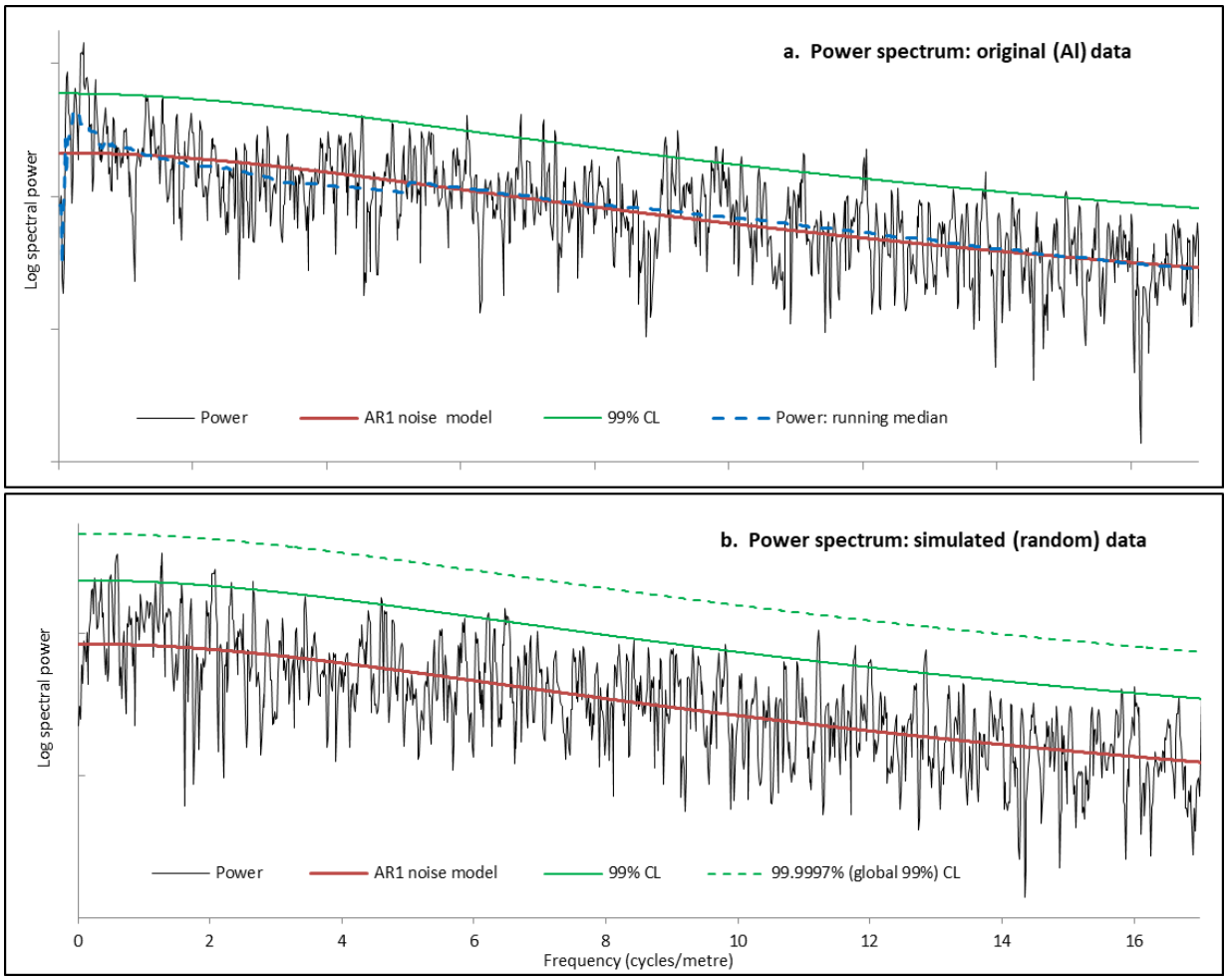

Fig. 1 Power spectral analyses of original and simulated (random) data-series of Zhao et al.[1], using the method of Mann & Lees (ML96)[5]. Compare Zhao et al.'s Supplementary Fig. 4a. The 99% confidence limit (CL) was used by Zhao et al. as their primary means of detecting candidate orbital frequencies in the power spectrum; here, it indicates comparable numbers of 'significant' frequencies in both the real and random data. Calculations were performed in Excel, and in R using the Astrochron software toolkit; for details see the Supplementary Information. Note that the linear frequency scale prevents illustration of the complete spectrum (which extends to the Nyquist frequency of 41.67 cycles/m).

a. Spectrum replicating Zhao et al. Supplementary Fig. 4a. Following their methods, the 73,721-point Al data-series (sampled at 1 mm spacing) was resampled to 6,141 points at 12 mm spacing, then detrended by subtracting an 8% LOESS weighted average. The method calculates the CL from

184 the ML96 'robust' noise model[5], which fits a first-order autoregressive (AR1) curve to the running
185 median of the MTM power spectrum.
186 **b**. Spectrum of 6,141-point (random) data-series simulated from the original running-median power
187 spectrum. MTM spectral analysis, robust noise estimation, and 99% CL were carried out as for Fig.
188 1a. A 99% 'global' significance threshold (i.e. applicable at all 3,000+ frequencies in the spectrum) is
189 the 99.9997% CL: it indicates no significant frequencies in this data-series which is, by definition,
190 aperiodic. This correction of the CL follows the Bonferroni method; other approaches are
191 available[8,9,10]. See Supplementary Information for computational details.
192