Statistics for cyclostratigraphic spectral analysis: what you need to know

David Smith
d.g.smith @ gmx.com

1. Introduction and objectives

This guide is for users of Confidence Limits on cyclostratigraphic power spectra, as generated using Mann & Lees (1996; ML96) and similar methods. It supplements a short presentation on the subject given at the Cyclostratigraphy Intercomparison Project workshop in Brussels (July 2024); I thank the workshop convenors for the opportunity, and for their support for the idea of this informal guide.

Confidence Limits (CLs) are a statistical tool; their appearance on ML96 plots indicates the existence of a statistical test of significance. Understanding how a basic significance test works is very important if the test is not to give unreliable results; fortunately it is not too difficult. Unfortunately, ML96 as used in cyclostratigraphy gives bad results *all the time*; one of the reasons is that the statistics underlying the CLs is largely concealed, especially when using one of the otherwise excellent software toolkits such as Astrochron or Acycle.

This guide provides a practical introduction to the workings of a Null Hypothesis Significance Test (NHST), and the relationship between an NHST and the scientific hypothesis for which it has been invoked.

My intention is that these notes could be consulted:

- If you are thinking of using ML96;
- If you have used it and want to understand or re-assess its results;
- If you are doubtful about someone else's work.

I also hope that these notes might form the basis for establishing some common ground among cyclostratigraphers who are concerned to understand the statistics better, and to avoid questionable results.

I am a geologist/cyclostratigrapher, not a statistician, and one of my recommendations is that you should consult a statistician if any of this material is not clear. But, I would be glad to help if I can; just send me an email.

David Smith
d.g.smith @ gmx.com

2. Introduction to Null Hypothesis Significance Tests

Statistics is a mathematical tool that can be invoked to test numerical observations from which the existence of some effect or pattern has been proposed. Statistics can test the probability that the observations are random; if they are not random then the proposed effect may indeed be real. Statistics has very strict rules of conduct, and it is important that its calculations should not be interfered with. Statistical calculations must be done under their own rules, and not directed by scientific expectations.

In order to keep the statistics distinct from the science, it can be useful to think of the Scientist and the Statistician as two separate people, thus:

> The Scientist explores some data, suspecting the existence of a pattern (such as cyclicity); the Scientist proposes this as a (Scientific) Hypothesis. If this hypothesis can be expressed numerically, it can be passed to a Statistician for a test of significance.
>
> The Statistician takes the Scientific Hypothesis and expresses it numerically, as a Null Hypothesis. The NH is so-called because it expresses the opposite point of view, that the observed data are effectively random. The Statistician tests the NH against the data; if the test cannot reject the Null Hypothesis, then the NH is accepted, implying that the observations have indeed arisen by chance. This result – passed back to the Scientist – is that the Null Hypothesis Significance Test does not support the existence of the proposed effect.

It is important to note that *the Statistician does not directly test the scientific hypothesis*. Statistics can only supply the probability that some observed value is effectively random. Thus, in cyclostratigraphy, statistical tests don't directly answer questions about the existence of orbital cycle-periods in stratigraphic data; statistics can only tell us about certain numerical properties of an observed value in a dataset.

3. Using NHSTs in cyclostratigraphy: the ML96 plot

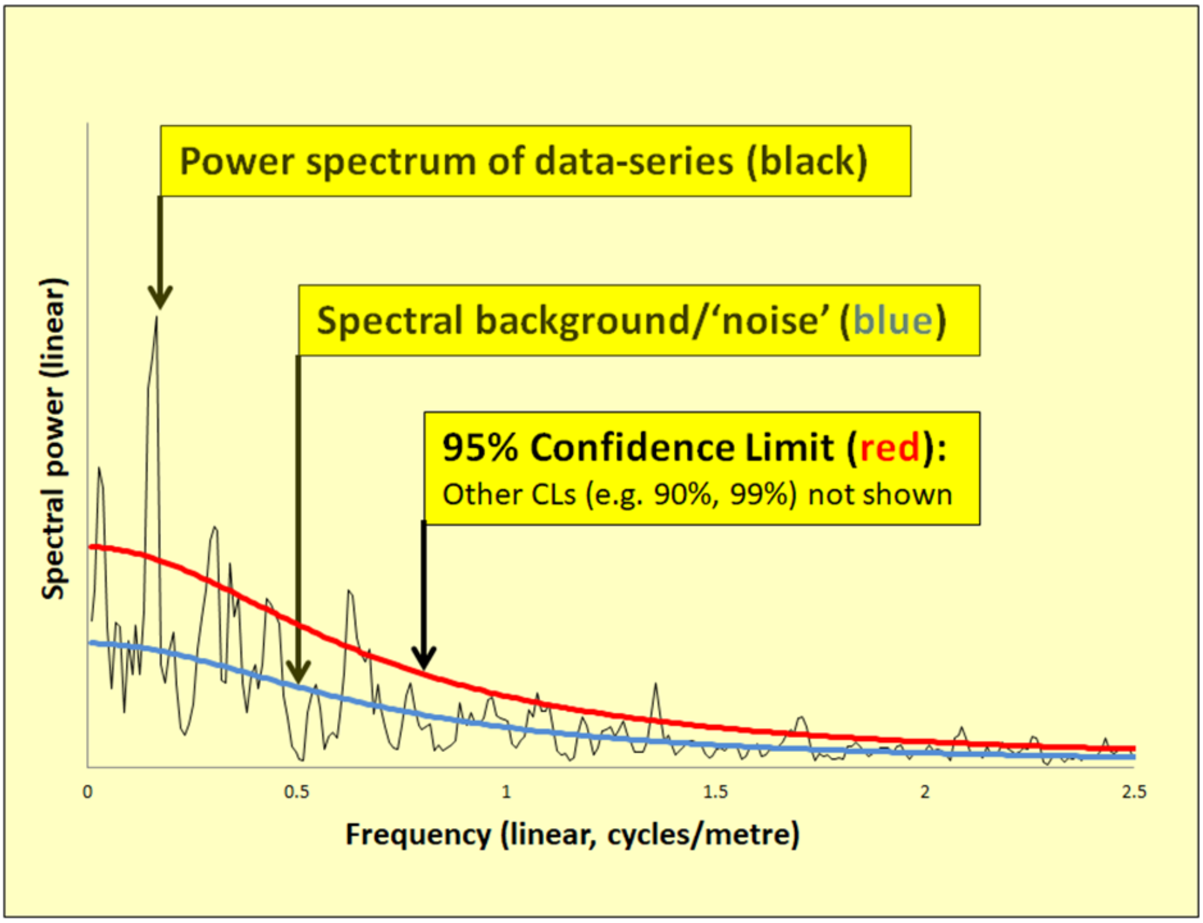

FIGURE 1: *The ML96 Plot: example of power spectrum with spectral background (often called the Noise Model) and one Confidence Limit. Spectrum, Spectral background, and the CL were generated (simultaneously) from a 1000-point data-series, using Astrochron's mtmML96 function.*

In cyclostratigraphy, the result of a spectral analysis is generally displayed with Confidence Limits (CLs), and the 'Noise Model' from which the CLs are derived. The widely used method of Mann & Lees (1996 – ML96) has been borrowed from climate science, where it was created to find the 'hockey-stick' trend in global temperature data. (It is important to note that ML96 was not intended to be used to search for significant peaks in cyclostratigraphic power spectra. ML96 is actually quite unsuitable for use in cyclostratigraphy, unless it is used with important modifications.)

Figure 1 shows a typical plot of a cyclostratigraphic analysis using the ML96 procedure. The power spectrum (in black) is accompanied by a noise model (in blue) and at least one Confidence Limit (in red; only a 95% CL is shown here, for clarity.)

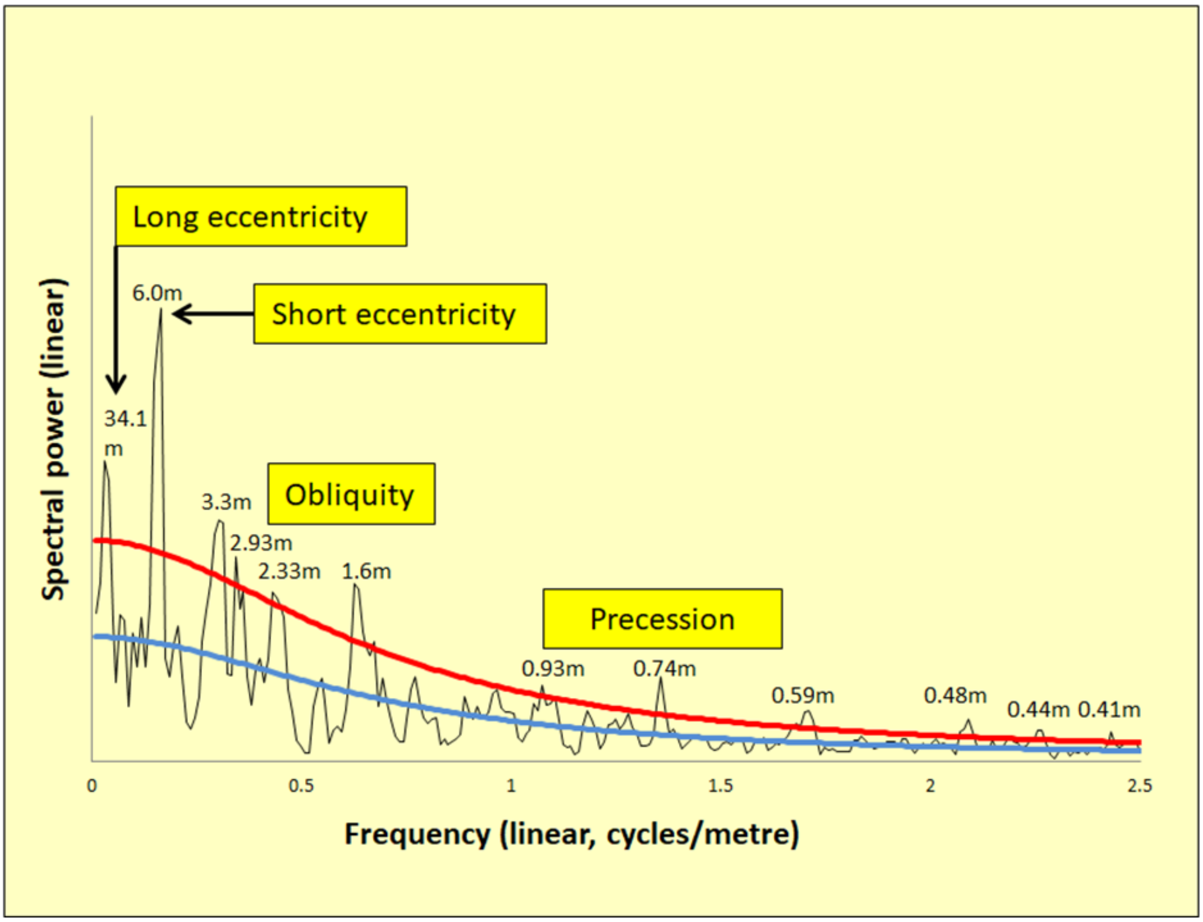

FIGURE 2: *In this interpretation of Figure 1; the 95% CL has been used to pick spectral peaks of potential interest. A template based on predicted orbital ratios has been added, to propose identification of peaks (or groups of peaks) with the principal orbital cycle-periods.*

Figure 2 shows the conventional use of the ML96 plot in a cyclostratigraphic investigation. The confidence limits are used as a guide to the detection of significant power peaks, and hence to the identification of candidate cyclic wavelengths. Other criteria – in this case, orbital ratios – have been applied to finalise the selection of peaks of interest.

4. Hypothesis - what hypothesis? If ML96 CLs represent a NHST, then what is it?

The 95% Confidence Limit (CL) in Figs 1 and 2 clearly represents a statistical test; in fact, a Confidence Limit ***is*** a Null Hypothesis Significance Test. In which case, what Null Hypothesis is it testing, and what Scientific Hypothesis does this NH represent?

It is a central weakness of the ML96 procedure that this essential hypothesising stage is omitted. Software implementations of the method do not require the user to pose any kind of hypothesis, and the impression that CLs are somehow independent of a formal NHST is reinforced by the various software implementations of ML96 (SSA-MTM, Acycle, Astrochron, …).

As a result, the graphical presentation (the ML96 plot) seems to imply that confidence limits emerge directly from the data, unconnected to any specific test. In fact, CLs do not exist in the absence of a

Null Hypothesis. Therefore, the CLs that appear on ML96 plots must represent a specific Null Hypothesis, but what is it?

Using Figure 1 as an example, we can reconstruct the Null Hypothesis from the way the ML96 plot is used in cyclostratigraphy. The plot is used to search for frequencies at which power exceeds one or more CLs; the CLs denote statistical significance at the given level of confidence, in this case, 95%. At this point, you should be asking the question: 'the statistical significance *of what?*'.)

So, the reconstructed Null Hypothesis for Figure 1 is this: ***Spectral power at every frequency is random***. The 95% CL suggests that this NH can be rejected at all frequencies where Power > 95%CL. For all of the 60 ML96 plots investigated by Smith (2023), this is the *implied* Null Hypothesis.

This Null Hypothesis is the Statistician's representation of the cyclostratigrapher's objective, which is to search the spectrum for candidate orbital frequencies. But the ML96 procedure bypasses the requirement for the Scientist to specify a testable hypothesis; it also bypasses the statistical step of representing this in probabilistic terms, as a Null Hypothesis for acceptance or rejection.

The result of obscuring this hypothesising stage is that ML96 *never* tests the right hypothesis! ML96 also obscures another feature of the NHST, which is the existence of many *Confidence Intervals,* collectively represented by the conventional Confidence Limit. This takes us into the inner workings of an individual significance test.

5. The workings of a Null Hypothesis Significance Test (NHST)

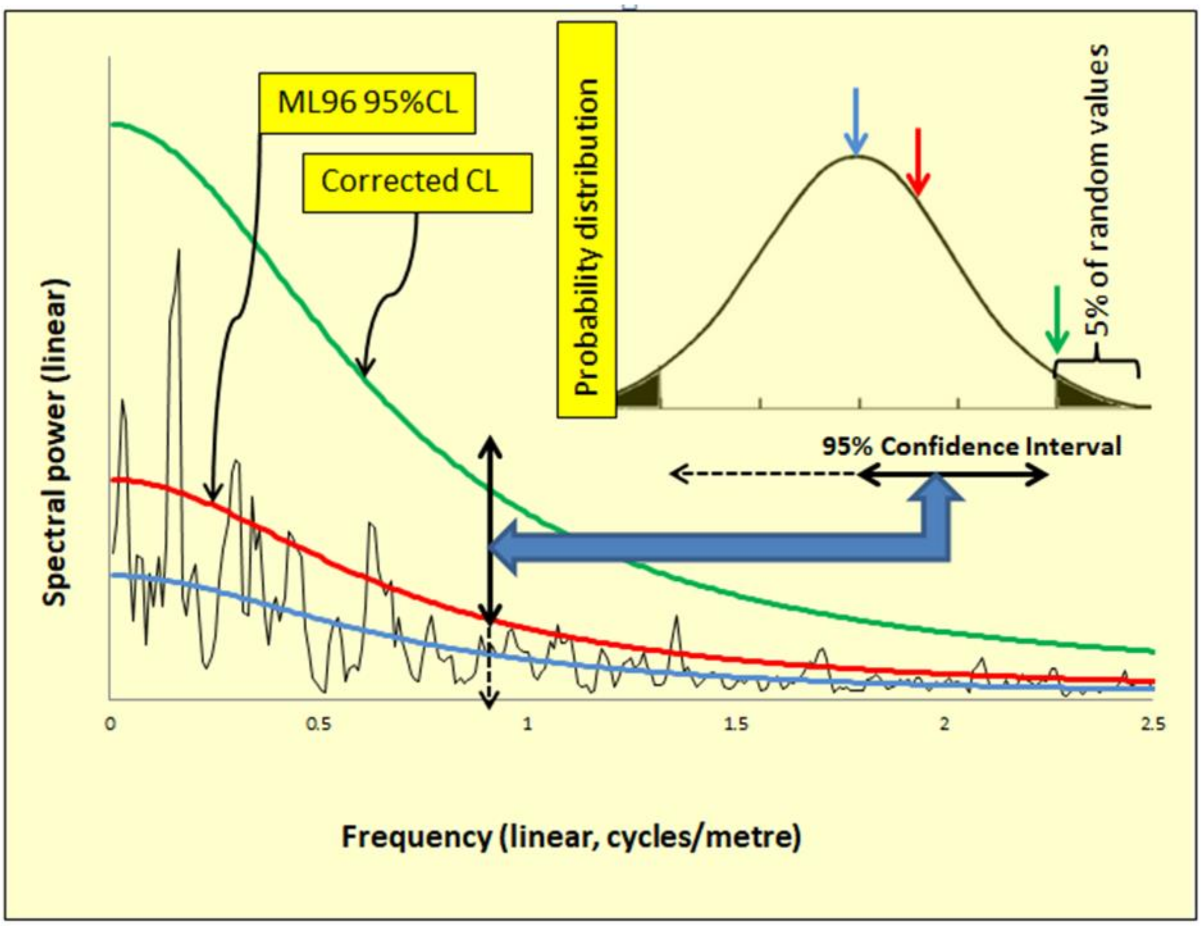

FIGURE 3: *Detail of the significance test [inset] at one frequency (double-headed black arrow). In a search for significant peaks, there are 500 such tests (in this particular example), and 500 such probability distribution curves, each presenting the 95% confidence* interval *at that frequency.*

The blue arrow shows where the median estimated spectral background intersects this curve; the red and green arrows show the intersections of (red) an uncorrected (ML96 default) 95% CL, and (green) a 95% CL corrected for multiple-testing at 500 frequencies.

IMPORTANT NOTE: In cyclostratigraphy we are interested only in the probabilities associated with unusually high values of spectral power; we are not interested in the part of the confidence interval that lies below the noise median (dotted arrow). The sampling calculations that give us the CIs at 95%, 99% etc take this into account.

It should be clear that far more than 5% of random power values exceed the uncorrected 95% CL, which is why it identifies substantial numbers of false positives as significant.

A Null Hypothesis Significance Test estimates the probability that an observed value is a chance pick from a random distribution. (It cannot be over-emphasised that this is all it does; the NHST does not itself answer scientific questions.)

In cyclostratigraphy it is the (many) observed values of spectral power that are to be tested for significance – are they likely to be real, or do those 'peak' values simply represent the more extreme end of a normally distributed random variable?

For a single observation, we need the probability that the observed value lies within, say, 2 standard deviations from some central value (such as the mean or median); this is called the 'Confidence *Interval*' – for an example, see the inset to Figure 3.

The height of the Confidence Interval depends on information about the possible distribution of random values, so we need the following:

1. the shape of the distribution; this is often assumed to be 'normal' (i.e. a bell, or Gaussian curve, like the example in Figure 3);
2. a central value – a baseline from which to measure the departure of the observed value;
3. the percentage (e.g. 95%) of the distribution that is to lie within the Confidence Interval.

To use the test, if the observed value lies within the Confidence Interval, then the NH is accepted, meaning that the observed value of power is random (within the parameters of this test). If the observed value lies outside the CI, the NH is rejected, with a probability $p < 0.05$: the test's conclusion is that the observed value is not random.

The job of the Statistician is to define and execute this probability test, and to report the result back to the Scientist. The statistical test does *not* answer (or even address) the scientific question; it provides a statement of probabilities (which may or may not be what the Scientist wanted to hear).

It should now be clear that the familiar Confidence *Limits* generated by ML96 are lines connecting the tops of the Confidence *Intervals* at every frequency in the spectrum. Now we must address the problem that the statistical probabilities for testing at *every* frequency are not the same as the probabilities for testing at *a single* frequency.

6. A problem: ML96 provides CLs for the wrong test

Now for a disclosure: The dataset represented in Figure 1 is *not* cyclic; it was based on random numbers, therefore *none* of the spectral peaks is statistically significant. This is because the Confidence Limits are too low. This is true of ALL cases where ML96 has been used without modification, and the simple reason is the arithmetic of probabilities.

The existence of this problem was well known to Mann & Lees (1996; see the very last sentence of their paper). For them, in their search for a *trend* in their data, it was less important. In cyclostratigraphy's search for *cycle-periods*, it is disastrous. The conventional procedure does not make it clear that the default Confidence Limits on ML96 plots are correct only for a single significance test, at a single frequency.

The apparently significant peaks in Figure 2 are *all* due to this problem: the 95% Confidence Limit is in the wrong place for this *multiple* test of significance. (A corrected 95% CL, for testing at all frequencies, is shown in Figure 3 (green line); it correctly identifies none of the spectral peaks as significant.)

The CLs in ML96 plots represent a NHST, but – for reasons outlined above – it is not clear that ML96 does not offer the test that we want. Whereas we want to *search* the spectrum (across all frequencies) for statistically significant peaks, the default NH applies only to a single frequency.

If we repeat a single-frequency test at all frequencies (there are 500 in Fig. 1), ***we change the probabilities***. The probability of a False Positive is repeated (in Fig. 1) 500 times, so we can expect about 25 false positives (500 x 0.05 = 25). The number of frequencies where power > CL in Fig.1 is consistent with this prediction.

7. Another problem: ML96 tries to fit the wrong spectral background

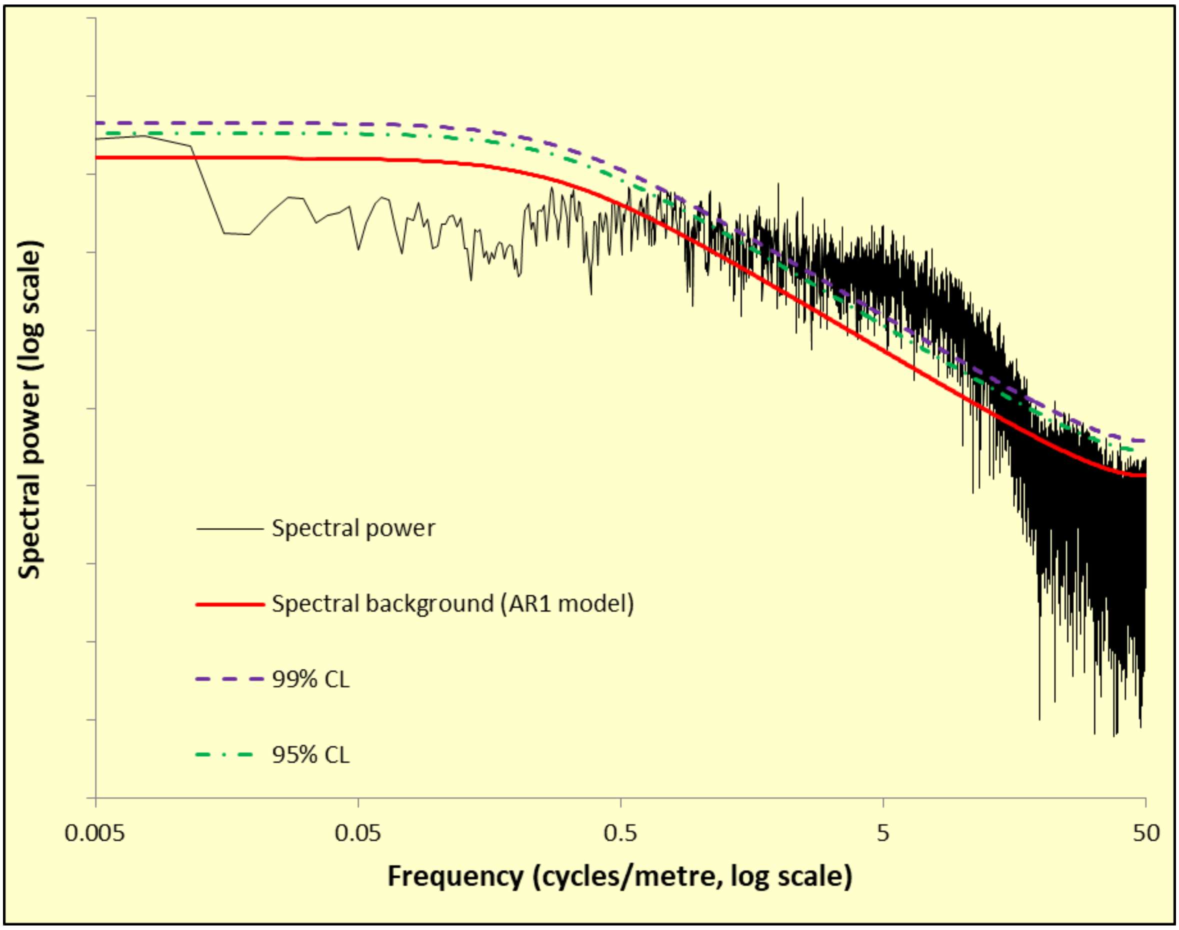

FIGURE 4A. *The importance of basing the spectral background on the data's spectrum rather than on an arbitrary parametric model. Power spectrum, log-log plot. MTM spectrum, 'noise model' (red), and two confidence limits, calculated (simultaneously) with Astrochron's mtmML96 function. Compare with Figure 3: the central values for the estimated noise distribution (the 'bell curves' at all frequencies) are far too high in much of the lower part of the frequency range, too low in the middle, and too high again for the highest frequencies. Significance tests relying on this estimation of the spectral background will be completely meaningless. (Figure based on Smith 2023, Fig. 3A.)*

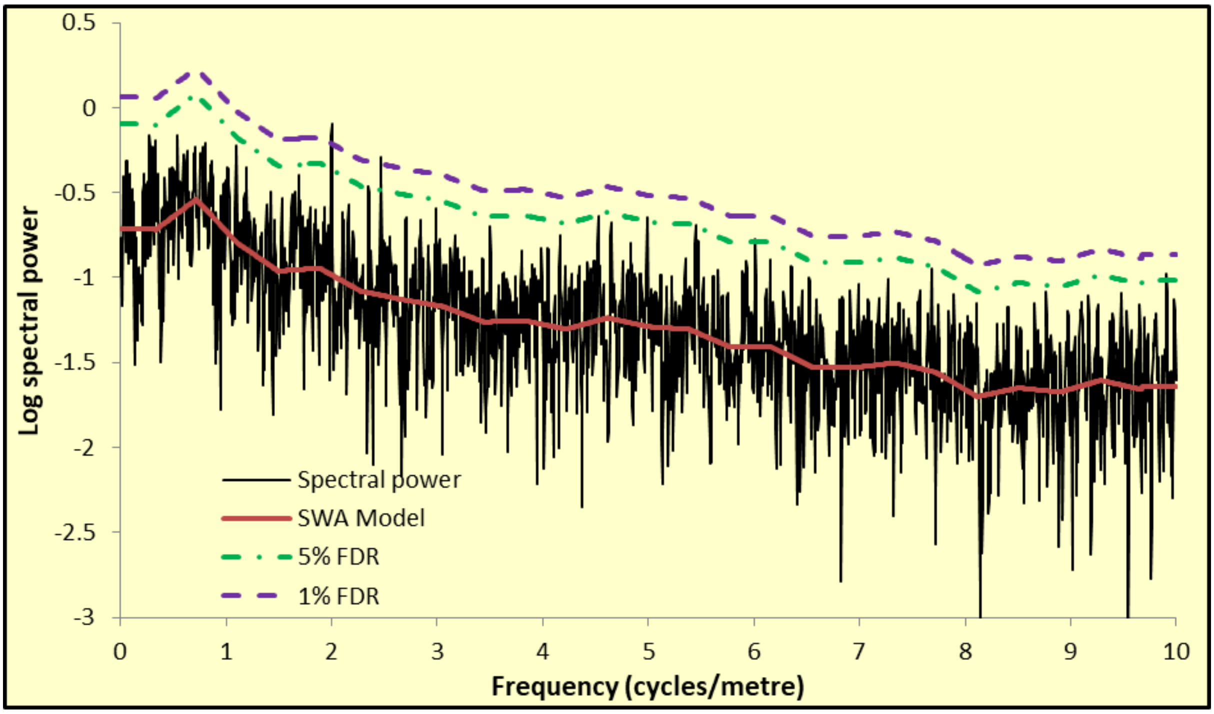

FIGURE 4B. *Same dataset (though with reduced number of points) plotted on linear axes. Spectral background estimation follows the Smoothed Weighted Average method of Weedon (2020, 2022). The visual fit between the spectral background and the data's power spectrum is now much closer; the central values of the ca. 2,500 confidence intervals (cf. Figure 3) will now provide realistic probability ranges for False Detection assessments at all frequencies. (FDR (False Detection Rate) is here preferred to 'Confidence Limit', for reasons explained below.) Based on Figure 4 in Smith (2023, q.v. for minor modifications to the data and the SWA method).*

The central value for the model of noise should provide a valid baseline for assessing probabilities: it defines the maximum point of the (half-)normal curve. (See the inset to Figure 3, noting again that only the upper half of the normal curve is of interest to us here.)

In Figure 4A, as in many ML96 examples, the default ML96 noise model is a very obviously poor fit, and this is especially evident on log-log axes. (See Table 1, column 9 in Smith 2023 for many more examples). A big part of the problem is the conventional insistence that the model must conform to AR1: (a) there is no reason for this (other than adherence to convention); and (b) it most often provides a poor fit.

The *only* source of information for estimating the spectral background is the spectrum of the data: it follows that the spectral background must be non-parametric (and hence free of preconceptions).

Although the need for a change in conventional practice is beyond dispute, methods for estimating the spectral background are open to discussion; Weedon's SWA is a very good starting point (see Figure 4).

NOTE: 'Spectral Background' is a much better name than 'Noise Model' to describe the numerical representation of all the spectral information that is *not* cyclic. The term 'Noise Model' is more frequently used in cyclostratigraphy, partly because the practice of forcing the background to fit an

AR1 curve is indeed a modelling process. Given that it's better to base the background *empirically* (i.e. on the actual power spectral values), it is not appropriate to call it a 'model'.

8. More problems: Interfering with statistical tests: Don't move the goalposts!

Statistics is arithmetic: statistical tests invoke the arithmetic of probabilities. Statistics, like regular arithmetic, has rules: once the procedure for a test has been defined, it must be followed exactly. Any departure from the pre-defined procedure represents a different test (with different probabilities, and different outcomes).

If we interfere with regular arithmetic and change the rules in the middle of a calculation, we get bad answers (e.g. 2 + 2 = 7).

If we interfere with statistical tests, we get similarly bad answers, because such interference ***changes the probabilities***.

The multi-test problem is an example: the default test is fine, but only for a single-frequency test; if we use it multiple times it will give meaningless answers. (The use of bad statistical practice to ensure the appearance of results is known as ***p-hacking***).

Two more ways of interfering with the validity of a test are Procedural Flexibility, and HARKing.

9. Procedural flexibility

Analytical procedures in cyclostratigraphy are widely regarded as flexible. Although basic procedures are usually similar, cyclostratigraphers are in the habit of varying the details to suit individual datasets. Such flexibility of approach is entirely appropriate in the early – exploratory – stages of analysis, but it is not compatible with the application of statistical tests, which operate correctly only under strictly controlled conditions.

Statisticians make a very useful distinction between ***Exploratory Data Analysis (EDA)***, and ***Confirmatory Data Analysis (CDA)***; this is the formal way of distinguishing between the work of the Scientist, and the subsequent, separate task of the Statistician.

For the Scientist's initial exploration of a dataset, the flexible approach is essential – any method that gets results is worth trying. This is the analytical stage that generates hypotheses to be passed to the Statistician for formal – confirmatory – testing.

For a Null Hypothesis Significance Test, flexibility leads directly to p-hacking: unreliable probabilities leading to meaningless results. This is because every procedural choice – every decision between two or more possible pathways – increases the *effective* number of significance tests: it ***changes the probabilities***.

Conventional analytical procedure in cyclostratigraphy follows a generally similar pathway, but does not specify the details in advance. Because ML96 does not demand an explicit hypothesis, the analysis is treated as exploratory, all the way from sampling strategy to interpretation of the power spectral plot. For example pre-processing steps may or may not comprise exclusion of outliers, resampling in various ways, de-trending, and filtering. Analytical steps include different choices of spectral methods and their exact parameters, and different approaches to estimation of the spectral background. Plotting of the results may be on linear or logarithmic axes, and may or may not include the entire spectrum; different confidence limits may be used in different studies. Flexibility is also applied to the selection of peaks among those indicated by the confidence limits. In EDA, you can try out all the different possibilities; in CDA, this isn't possible without ***changing the probabilities***.

The problem with procedural flexibility is that the statistics of the NHST cannot cope with the potentially huge number of scenarios that arise from this 'Garden of Forking Paths'. If you throw enough dice, you will always get a six. But this result does *not* tell you anything about the probability that an individual observed result has arisen by chance, which is what you actually need to know.

10. HARKing: Hypothesising After the Result Is Known

P-hacking is about trying many possibilities until you get the result you want. HARKing extends this to deciding on what result you want only after you have seen the analytical results. The classic analogy is the 'sharp-shooter' who fires his shots at a blank wall and only then draws his target – success is guaranteed! Again, this is fine in an exploratory, hypothesis-generating stage of analysis; it is not compatible with doing reliable confirmatory (statistical) tests.

Figure 2 exemplifies the conventional 'orbital ratio test', in which a template is fitted to a selection of 'significant' peaks. This inverts the proper EDA -> CDA sequence: the outcome of the ML96 analysis is explored as if it were primary data, whereas it should be treated as the product of a confirmatory NHST. Also, the number of possible templates is very large, and the statistical test already applied does not cover the implied range of scenarios.

'HARKing', or post-analytical target-setting, describes this procedure (HARKing stands for Hypothesising After the Result Is Known). The validity of a statistical test depends on the care with which it is designed; it also depends on strict adherence to a pre-defined procedure. Exploring different interpretations ***changes the probabilities***. You can't mix confirmatory NHSTs with an exploratory approach to its results.

In Figure 2 the final orbital interpretation is NOT what was being tested by the CLs; it is one of many possibilities. If we were to test that dataset for statistical significance at specific frequencies, the probabilities would be very different (because an exact match is so unlikely).

11. Prior Probabilities: adding commonsense to statistics

If ML96 is used (unmodified) on data from successions that cannot possibly record OFCC (a pile of basalt lavas, to take an extreme example), it ***will*** find statistically significant peaks in spectral power, because of ML96's uncorrected multiple use of significance thresholds. Cyclostratigraphy rarely takes any account of the *a priori* likelihood of cycles existing in a chosen succession, yet such a 'common-sense' approach is not only possible, but desirable.

If I (a male) take a pregnancy test that is 95% reliable, there is a 5% chance of my getting a positive result. Would that prove that I am pregnant? No, because the ***Prior Probability*** that I am pregnant is zero.

Similarly, if ML96 confidence limits find cyclicity in a succession of basalt lavas, it would be a good idea to consider the prior probability before jumping to any conclusions of orbital-forcing.

Common-sense suggests that some depositional environments are more likely than others to respond to orbitally-forced climate change, while many are very unlikely to record OFCC. Yet the unwritten expectation of all cyclostratigraphic investigations is that OFCC is present in the chosen dataset. Unfortunately, ML96 reinforces this expectation, by generating positive results regardless of how unlikely they are.

It would be a very useful exercise to assess the relative prior probabilities that OFCC will be recorded in various settings, and which orbitals can be expected at different paleolatitudes.

I conclude, below, with a list of suggested Do's and Don'ts, and with a short discussion of Types 1 and 2 statistical errors.

REFERENCES

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

Smith, D.G. 2020. Misplaced confidence: limits to statistical inference in cyclostratigraphy. *Bol. Geol. Min.* **131**, 291-307. doi: 10.21701/bolgeomin.131.2.005

Smith, D. G. 2023. The Orbital Cycle Factory: Sixty cyclostratigraphic spectra in need of re-evaluation. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 628, 111744. https://doi.org/10.1016/j.palaeo.2023.111744

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**, PA4211, doi: 10.1029/2011PA002195.

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. doi: 10.21701/bolgeomin.131.2.001

Weedon, G.P. 2022a. Cyclostratigraphy: regular cycles detected and counted to measure time. In: Coe, A. L. (ed.) 2022. *Deciphering Earth's History: the Practice of Stratigraphy*. Geological Society, London, Geoscience in Practice, 161–179. https://doi.org/10.1144/GIP1-2022-26

Weedon, G.P. 2022b. Problems with the current practice of spectral analysis in cyclostratigraphy: avoiding false detection of regular cyclicity. *Earth-Science Reviews* 235, 104261. https://doi.org/10.1016/j.earscirev.2022.104261

I can recommend two books on statistics, but both are in English, and it may be better to look for books in your own first language – ask your Statistics Department for suggestions.

Spiegelhalter, David, 2019. *The Art of Statistics: Learning from Data.* Penguin Random House, UK.

Vaughan, Simon, 2013. *Scientific inference: Learning from data*. Cambridge University Press.

A note on Type 1 and Type 2 statistical errors

In the debate about the validity of ML96 plots, counter-criticism often refers to Type 1 and 2 errors, the balance between them, and the relative (un)desirability of both.

A TYPE 1 ERROR occurs if a statistical test gives significance to an observation likely to be random; in cyclostratigraphy this results in a falsely significant peak in spectral power, potentially leading to incorrect identification of an orbital cycle period.

A TYPE 2 ERROR occurs if a statistical test fails to assign significance to an observation representing a real effect, wrongly classifying it as random. In cyclostratigraphy this would lead to failure to detect a real cycle-period in the data.

Clearly both types of error should be avoided, but the debate around them in cyclostratigraphy is largely irrelevant if the underlying methods are wrong.

An argument used by cyclostratigraphers is that a CL that is too high risks missing 'real' results, and that it's better to admit more peaks into significance because other criteria can be used to sort out real from random.

In Figure 1, ALL identifications of statistically significant observations are wrong, because the dataset is based on random numbers – there are no cyclic frequencies. In this case, any attempt to achieve a balance between Type 1 and Type 2 errors is irrelevant – the problem lies with the position of the CL.

It's a simple arithmetical fact that a multi-test (all-frequency) significance threshold is higher than that for a single-frequency test; this is a very simple consequence of the arithmetic of probabilities and nothing at all to do with T1 versus T2 errors. It is certainly NOT some conspiracy to reduce the number of frequencies at which significance is reached.

If the CL(s) is/are correctly placed, then ALL frequencies with Power > CL are significant *in terms of this test* – and selection among them is not possible. That is because any such selection implies a change to the terms of the test; it changes the scientific hypothesis, and hence it changes the statistical null hypothesis; and ***it changes the probabilities*** represented by the CL.

The Do's and Don'ts of statistical testing in cyclostratigraphic power spectra

Participants in the Cyclostratigraphy Intercomparison Project workshop in Brussels (July 2024) were asked to suggest Do's and Don'ts for statistical practice in cyclostratigraphy. These are my suggestions:

DON'T apply ML96 without reading about it and understanding it.

DON'T use ML96 plots simply because (1) everyone else uses them; (2) they are easy to use; and (3) they give lots of positive results

DO: Read up on the basics of statistical testing, and/or consult your university's Statistics Department. Understand that any action/factor that increases the *effective* number of experiments will change the probabilities. All such interference/changes *must* be compensated for by changing the statistical thresholds; flexibility of procedure, all-frequency searches of the spectrum, and post-analytical target-setting (HARKing) all increase the effective number of experiments, and therefore ***change the probabilities*** of the test statistics.

DO think of statistical tests as being separate from the rest of the scientific investigation. Any NHST must be valid on its own terms (or it isn't valid at all).

DON'T try to drive the statistics with 'special pleading', whether based on the perceived 'special' nature of stratigraphic data, or on over-optimistic expectations of what *should* be there.

DO consider what you want from an analysis, and think carefully: what is your Scientific Hypothesis? Can this be expressed numerically as a NHST: if it can't, don't use statistics.

If you decide to use a NHST, **DO u**nderstand why it is absolutely essential to make adjustments (read Weedon on this).

DO everything you can to reduce procedural flexibility: use a standard procedure; avoid pre-processing steps; base your estimate of the spectral background on the spectrum of the data, not on an arbitrary parametric model such as AR1; calculate CLs according to the number of points in the spectrum; etc etc.

DON'T allow expectations and optimism to drive your statistical tests. Plenty of convincing cases of orbital cyclicity exist for which power spectral results were inconclusive.

DON'T risk being accused of p-Hacking (also known as 'torturing the data until the desired results appear'.)

DO publish null/negative results if you can find a journal that will allow you to! Negative results are at least as important as positive results.

DO experiment with random datasets; they are easy to generate, and they tell you what 'random' spectra look like – essential if you are using statistics to tell real and random apart. Astrochron's makeNoise function generates synthetic data-series from any input spectral background; you can also generate simple random data-series in Excel.

DO consider replacing 'Confidence Limits' with False Alarm Levels/False Detection Rates (see the papers by Graham Weedon in the References).

DO try to change your mindset: you want to find cycles; a NHST wants you NOT to find cycles! Statistics is counter-intuitive; statistics is pulling in the direction of a NULL answer (Spiegelhalter describes the NH as *relentlessly negative*). It's too easy to invalidate a statistical test by trying to pull the statistics in a positive direction by p-hacking, HARKing etc etc.

DO please abandon the conviction that ML96 gives usable results. Do engage in discussion about ways to modify it; better still, discuss ways of replacing it altogether.

DO be aware that statistics is a technical tool; understand it, follow its rules, and respect its results.