The Garden of Forking Paths: the Hidden Statistical Consequences of Data Contingency and Researcher Degrees of Freedom in Cyclostratigraphic Analysis: Why Most Published Results are False

David Smith* Independent consultant, Truro, U.K.

The problem – not unique to cyclostratigraphy – is STATISTICAL MULTIPLICITY:
"repeated looks at a data set in different ways, until something statistically significant emerges"
www.statistics.com

Why is MULTIPLICITY relevant to cyclostratigraphy? **It's all about what is being asked of the data …**

The cyclostratigrapher's question is:

Are there any cycles, if so, how many, and at what frequencies?

This is *not* a statistical question, but is typical of …

EXPLORATORY DATA ANALYSIS (EDA):
- Searching for patterns (e.g. cyclicity) in order to erect hypotheses:
- Use of multiple techniques and parameter values is essential

EXAMPLES: Figures 1 and 6

Whereas a typical question for statistics is:

Could a spectral peak at frequency F be due to chance?

This is a *strictly* statistical question, and is central to …

CONFIRMATORY DATA ANALYSIS (CDA):
- Testing a hypothesis for statistical significance:
- Strict protocols are critical; no flexibility; accept/reject hypothesis

EXAMPLES: Figures 2 and 3

The conventional (incorrect) approach:
CONFIRMATORY testing used in **EXPLORATORY** mode

In this REAL DATASET[R16], where are the cycles?
No null hypothesis; uncorrected (black-box) Confidence Limits

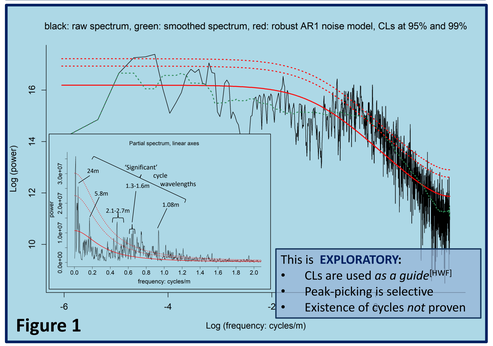

Figure 1

Data: Ca data (3522 values, detrended), Pliensbachian, Mochras-1, UK onshore, 861.4 - 1283.92 m, depth interval 0.12 m
Source, and cycle period/wavelength picks: Ruhl et al. 2016; see also Smith & Bailey 2018; Hinnov et al. 2018
Analysis: MTM (3 tapers), using unpublished R function mtmDS, modified from Astrochron function mtmML96 to plot on log-log axes

Compare these examples of CONFIRMATORY analysis:
each is a strict test of significance at a single frequency

1. Do monthly sunspot data show an 11-yr cycle period?
Statistical (null) hypothesis: Data are random at *f* = 0.091 cycles/year

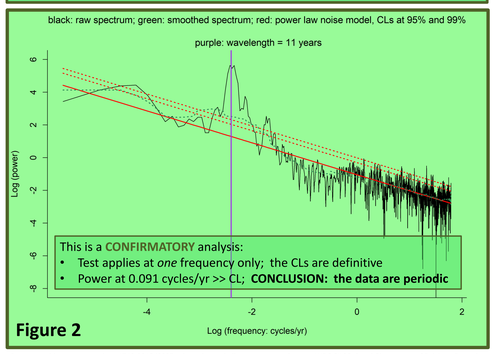

Figure 2

Data: monthly sunspot numbers 1749-2018 (3237 values, detrended) **Source**: WDC-SILSO, Royal Observatory of Belgium, Brussels
Analysis: MTM (3 tapers), using unpublished R function mtmPL, modified from mtmML96, to fit and apply a power law noise model

2. Is the English climate influenced by the sunspot cycle?
Statistical (null) hypothesis: Data are random at *f* = 0.091 cycles/year

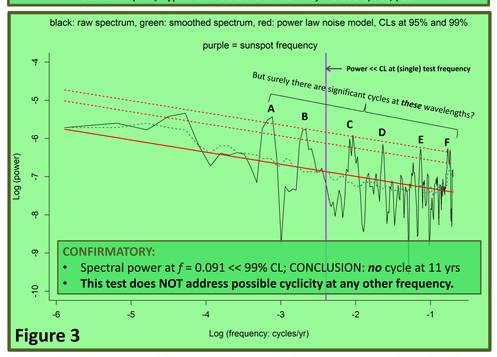

Figure 3

Data: Annual Central England Temperature (CET) record 1659-2018 **Source**: Met Office (UK) Hadley Centre for Climate Change
Analysis: MTM (3 tapers), using unpublished R function mtmPL, modified from mtmML96, to fit and apply a power law noise model

What about spectral peaks A to F?
'Testing' more peaks with these *single-test* CLs incurs
MULTIPLICITY, *which gets a result by throwing more DICE*

A GAME OF CHANCE:

Given the 1:6 probability of getting a six, why does nearly every throw include a six?

"**The more analyses you perform on a data set, the more your overall alpha** [false positive] **level increases.** Perform two tests and your chance of at least one of them coming out falsely significant is about 10%; run 40 tests, and the overall alpha [FP] level jumps to 87%. **This is … the problem of *multiplicity*, or *Type I error inflation*.**" [Pez.]

A CHALLENGE (1): make a throw of all 30 dice that does NOT include a Six.

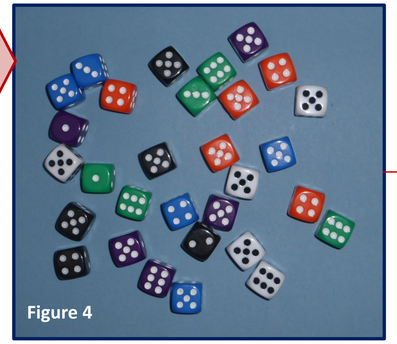

Figure 4

Unrecognised multiplicity leads to the wrong confidence estimates and to False Positive results (Type I statistical errors) .

Sources of multiplicity in cyclostratigraphy:

1. Assumed freedom of analytical method:
   - The **Garden of Forking Paths**[GL], a.k.a.
   - Researcher Degrees of Freedom[SNS]

   Figure 5
2. Single-test CLs used to *search* spectra[VBS]

   Figures 1 and 6

THE GARDEN OF FORKING PATHS

"Inadvertent multiplicity of analysis"[GL]: every data-contingent decision implies an alternative pathway

DATA	SECTION	Basin? locality / exposure / borehole? exact line of section?
	SAMPLING	Climate proxy/proxies (e.g. MagSus / CaCO3 / TOC / C-isotopes / GR / other …)? sample size and spacing?
METHOD	PRE-PROCESSING	Omit outliers? omit 'event' beds? re-sample to equal intervals: interpolate or downsample? de-trend? if so, linear or/and polynomial?
	PROCESSING	Which software package(s)? spectral analysis method? parameters?
	STATISTICAL TESTING	Choice of noise model (AR1 / PL / BPL / other …); model estimation method and parameters?
TARGET	CYCLE IDENTIFICATION	Plotting parameters: linear or logarithmic axes? full or partial spectrum? identification of 'significant' cycle periods: CLs binding, or a guide?
	TARGET	Selection of likely (combination of) orbital target cycles: variable sedimentation rate(s) (in ASM method)

Figure 5

For statistical purposes, the number of *cases* is the number of *potential* pathways.
As for the dice, the EXPECTED OUTCOME (hence the P-value) is conditioned by the number of cases.

The more dice that are thrown, the greater the inevitability of a positive outcome

A CHALLENGE (2): analyse a *random* dataset in the conventional way (e.g. ML96) WITHOUT finding 'significant' frequencies ….

The conventional (incorrect) approach finds 'significant' cycles in RANDOM data[VBS]

Conventional analysis: RANDOM DATA

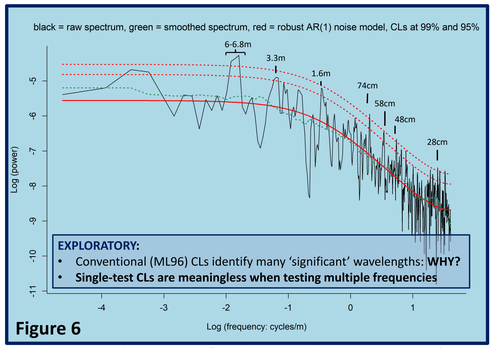

Figure 6

Data: 1024-pt random dataset, 1st-order autocorrelated, coefficient = 0.7, sample interval = 0.1m. **Source**: Astrochron function ar1
Analysis: MTM (3 tapers), using mtmDS, modified from Astrochron function mtmML96 to plot on log-log axes. Spectrum has 512 points.

Exploratory spectrum search *is* possible, but only if CLs are corrected for testing at multiple frequencies[VBS]:
- The above search implies tests at *N*/2 = 512 frequencies (*N* = 1024)
- To correct CLs, divide α (the False Positive rate) by 512 (α + CL = 1)
- For a *global* 95% CL, α = 0.05/512; corrected *local* CL = 99.99%

Contrary to critical comments[HWF etc], this correction is neither 'unrealistic' nor 'extreme': uncorrected CLs may appear to give a desirable *cyclostratigraphic* outcome, but at the expense of any *statistical* integrity. For real data, further corrections should be made, to account for data-contingent analytical multiplicity[G].

RANDOM DATA: CLs corrected for whole-spectrum search

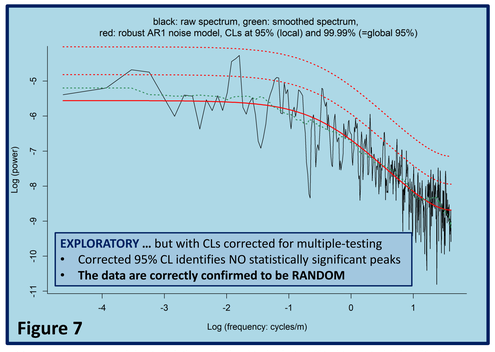

Figure 7

Data: As Figure 7. **Analysis**: mtmDS, further modified from Astrochron's mtmML96 to plot a non-standard, 'global' Confidence Limit.

Where does this leave us?

The need for multi-frequency CL corrections has recently been accepted (Meyers 2018);
The Forking Paths route to Multiplicity remains unacknowledged and is more serious;
Statistical methods (and results) in cyclostratigraphy urgently need a full review.
If in any doubt, ***ASK A STATISTICIAN!***

Comments, please …

***d.g.smith@talktalk.net**

References
and recommended reading

Benjamin, D.J. and ~50 others, 2018. Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6.

Button, K.S., Ioannidis, J.P., and others, 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Revs Neurosci.* *14*(5), 365.

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

[GL] Gelman, A. and Loken, E., 2013. The garden of forking paths: Why multiple comparisons can be a problem, … *Department of Statistics, Columbia University*.

[HWF] Hinnov, L.A., Wu, H. and Fang, Q., 2016. Reply to the comment on "Geologic evidence for chaotic behavior … *Palaeo3*, *461*, pp.475-480.

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

[Mey] Meyers, S. (in press). Cyclostratigraphy … astrochronologic testing. E.Sci.Rev.

[Pez] Pezzullo, John, How to Handle Multiplicity in Clinical Trial Data, dummies.com

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 News*, *526* (7572), 182.

[R16] Ruhl, M. and others, 2016. Astronomical constraints on the duration of the Early Jurassic Pliensbachian Stage and global climatic fluctuations. *Earth and Planetary Science Letters*, *455*, pp.149-165. *Comment/Reply*, EPSL 481, 412-419.

[SNS] 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), pp.1359-1366.

Stark, P.B. and Saltelli, A., 2018. Cargo-cult statistics and scientific crisis. *Significance*, *15*(4), 40-43.

[VBS] 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*(4).