Nonlin. Processes Geophys. Discuss.,
https://doi.org/10.5194/npg-2018-38-RC2, 2018

Nonlinear Processes in Geophysics
Discussions

Interactive comment on "Exploring the effects of missing data on the estimation of fractal and multifractal parameters based on bootstrap method" *by* Xin Gao and Xuan Wang

Anonymous Referee #2

Received and published: 23 December 2018

The article treats a very interesting subject and distinguishes itself by comparing a wide range of interpolation methods. The subject is vast and an exhaustive treatment is not trivial and I commend the authors for their work.

On the face of the results, it appears to me none of the interpolation methods clearly outperform the basic method of using "series containing missing values". Given that the method, which appears to be the same as detrended fluctuation analysis (DFA), can accommodate missing data, I would see the pure randomly gliding boxes method as the clear winner over the interpolation methods. This is interesting as it confirms the intuition that when interpolation can be avoided, it should. I think such studies would

be more interesting to perform with methods such periodogram or multi-tapers which usually require uniform data and thus interpolation, rather than with DFA which does not require interpolation.

However, I cannot recommend acceptance of the paper at the moment for reasons of language, order and content. At best, there should need to be major revisions made in both the writing and the numerical experiments, but given the large amount of changes I believe it requires, I think that it will be easier for the authors to resubmit the paper anew. Below I give evidence of the need for major revisions.

Language

I started making detailed comments on the improvements needed in the writing of sentences, but then stopped writing them all since most sentences need to be revised. Here is the beginning:

Line 53: Define the following acronyms: PLI and PBI

Line 55: Replace 'are unable to ignore' by 'cannot be ignored'

Line 82: I think you did not mean to have '=' there in '0<alpha=<1'

Line 164: I fail to see a difference between regression residual variance and detrended fluctuations analysis (DFA). Did you mean DFA?

Line 163-176: It seems to me that the explanation of the regression residual variance is misplaced, right in the middle of the explanation about fractional Brownian motion.

Line 195: "the" scaling function (an article is needed in front of scaling)

Line 197: "a" universal generalized (u is pronounced like a consonant here, therefore "a" rather than "an".)

Line 197: Schertzer and Lovejoy missing from reference list at the end.

Line 200: Replace "of which have" by "used are"

Line 201: Replace “involving lots” by “it involves a lot”

Line 242-245: Sentence is convoluted, needs rewrite.

Line 246-248: Sentence is convoluted, needs rewrite.

Line 247: On the condition that it is poorly known? I think you mean “even when it is poorly known”.

Line 255-258: Sentence is convoluted, needs rewrite.

Line 260: “By the glide of a group of boxes”? I think “using a group of gliding boxes” is what you mean.

Line 262-264: Sentence is convoluted, needs rewrite.

Order

Then, another part of the writing is what I will call here order, and in which I mostly include problems of definitions. The symbols used are rarely defined properly (F, G, Theta, sigma, s, c, etc). This ties in with problems of structure. The biggest problem I identified is the lack of numbered equation. Any equation which is important warrants a number. I think a good way to go about writing the theory section, is to introduce first an equation, and then define in the most straightforward manner, what the terms in the equations stand for, and then give explanations of the significance of the terms and the equation. Below, I give a few points as proof again that major revisions are needed in this aspect.

Line 204: I don’t think int() is a standard function, you should indicate that it means you are taking the integer part of the fraction.

Line 278: What is meant by concealment?

Line 278: If there is such “concealment” to produce Q, then Q isn’t really a population anymore, but rather a sample right?

Line 327: What is meant by standard deviation of distribution with double sampling?

Line 324, 328: R and S are the re-sampling and secondary sampling, but I do not think these were defined clearly. It would help to indicate on the diagrams when the sampling R and S happen. On this note, I am unsure of the reason for the necessity of doing this resampling twice. I think it would be good to explain this better.

Fig 3, 4: Captions should indicate directly which is (a) and which is (b) without the need to refer to the text.

Fig 14: It does not make sense to test for C1=0.9 here. The range of tests could be restricted under 0.5, and the resolution increased to maybe 0.05. Given the unsmoothness and lack of monotonicity, it appears that the number of replicas in the sample might not be sufficient, or at least, I do not see how it makes sense that there is a bump at 0.4.

Line 392: Why is it decreasing? Is that a feature one should expect? Could it be random because the variance of the estimator becomes increasingly large?

Line 427: What is this “power function”. I do not think it was properly introduced, and the usage made of it remains vague.

All tables: The numerical results should be given with consistent significant figures

Line 569: Again, giving 1.992 seems overly precise. Generally, the number of sig figs corresponds to the presumed accuracy of the estimates. Unless the accuracy can be to three significant figures, which is not the case here since the error seems to be on the order of 0.1, then it should be rounded to 1 sig figs, i.e. this result could be reported as 2.0 +/- 0.1 and the mean estimate would be virtually equal to a log-normal distribution.

Content

The numerical experiments could be greatly expanded. The most ominous point is that only one set of parameters if tested for each type of series, i.e. fBm and multifractals.

The authors should be aware that the results and conclusions they make, might very well be dependant on the value of those parameters, and I expect they are. So while one type of interpolation method might perform well for H=0.6, another might perform better for H=0.3, and so on. Same applies for C1 and alpha. Therefore, for the paper to be more relevant, I would recommend expanding the experiment to all values of H between 0.1 and 0.9, and all values of C1 between 0 and 0.3, and all values of alpha between 1. and 2. Also, in addition to the RMS value, I think the authors should consider reporting the bias and variance of the estimators. Of course, the authors will realize when they do that, that the vastness of results produced is challenging to report, and that new methods of visualization might be necessary. I believe this will also greatly affect their outlook on their conclusion and the significance, and what is important to report.

Interactive comment on Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-38, 2018.