ELECTRONOTES
1016 Hanshaw Road
Ithaca, NY 14850

APPLICATION NOTE NO. 403

Jan 27, 2014

SPURIOUS CORRELATIONS DUE TO FILTERING (OF NOISE)

How disappointing. Something obvious has gone fairly unnoticed. Except that it was noticed at least 90 years ago and now apparently has a name which is new to me: "Slutsky/Yuel" [1]. All it is, so it seems to me, is filtered noise. We know that when we filter white noise, it becomes colored; the spectrum is no longer flat. It will become correlated in some manner. We do this a lot. So what's the big deal? Well – what if you do something that is really filtering but you don't make a big deal of it? Is it still filtering? Of course it is. You can end up with spurious correlation.

Often we have data, and we want to process this data, or at least look at in a different or more comfortable way. Accordingly we do things like remove the trend (detrend), which is a DC (at least very low frequency) rejection, or we may smooth the data with a low-pass such as a moving average [2]. An eventual goal may be to look for a correlation between two data sets. No one should make the mistake of discovering correlation and assuming causation, or even that the data sets are jointly caused by something else. But correlation can still be an essential step, because if you don't have even a correlation, you have nothing. But correlation could be spurious, or an artifact.

So at what point in your data processing/presentation should you try your correlation? On the original nasty data? Or on the cleaned up data? Does it even matter? Yes, because filtering, even if just to neaten thing up, can cause spurious correlation. This is kind of obvious after someone reminds you of it. Here we will begin with two completely random white (uncorrelated) noise sequences and show how you can get correlation.

We might suspect that our noisy data has a low-frequency "trend" with a bunch of higher frequency wiggles that interest us. (Or, it might well be the other way around.) In such a case, we might want a detrending operation. This we could achieve by subtracting successive samples (essentially differentiation) or subtracting samples separated by a particular sample interval. This can reduce a low-frequency component. Possibly more familiar is the moving average that does the opposite. We often think of it as removing "noisy" higher frequency "fluctuations" which we assume have little meaning. We are somewhat unlimited in what we assume will be useful. Here we note that these operations are well-known FIR digital filters.

Fig. 1

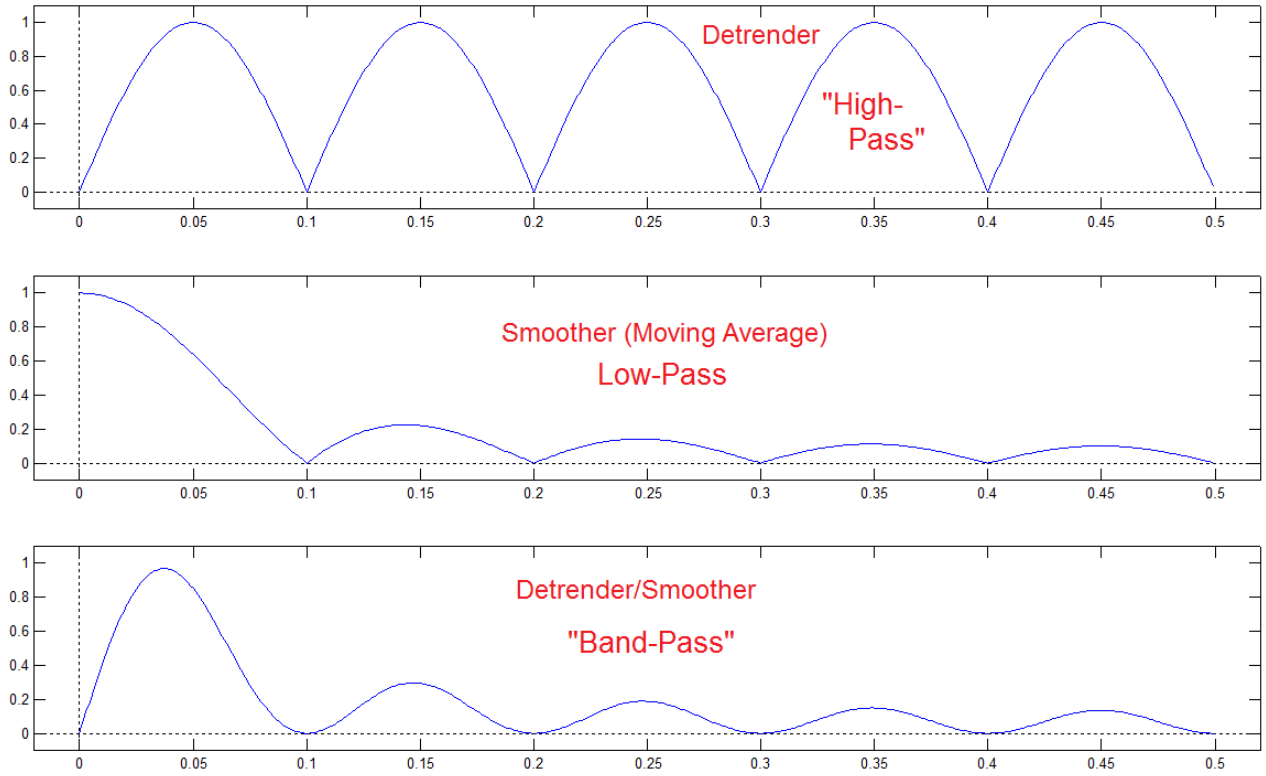


Fig. 1 shows three filers we will use in our demonstration. The detrender has impulse response:

h1 = (1/2) [1 0 0 0 0 0 0 0 0 0 -1]

while the Smoother is a standard length-10 moving average:

h2 = (1/10) [1 1 1 1 1 1 1 1 1 1]

while the Detrender/Smoother combines both (convolves impulse responses) as:

h3 = (1/15) [1 1 1 1 1 1 1 1 1 1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1]

where the constants out front is chosen to best use the display. The bottom panel of Fig. 1 is the multiplication of the top two panels. Note the second order zeros (flat nulls) at frequency multiples of 0.1 (but not at zero). Here the “sampling frequency” is chosen as 1. The bottom two panels are not that unlike, except at zero frequency. The middle panel is low-pass (a moving average) while the bottom begins more like a band-pass.

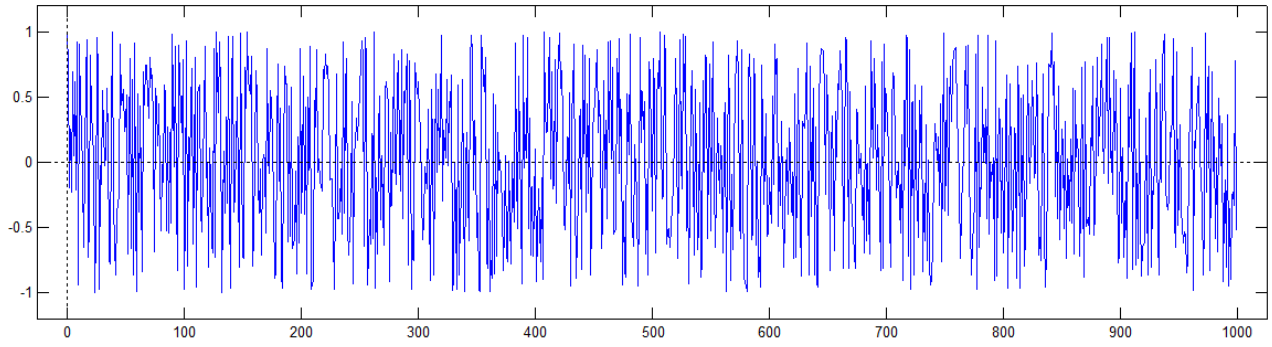
Keep in mind that we are not bragging about using filtering and (in consequence) shaping a spectrum. We are just improving the data by suppressing a lot of the junk. Well – so we say.

Paying particular attention to the region right above zero frequency, it is clear that the detrending operation is high pass while the smoothing operation is low-pass, although not great examples of these filters. We also remember that a series application of a high-pass and a low-pass can produce a band-pass in the overlapped region (multiplication in the frequency domain) as is seen in the bottom panel of Fig. 1. While any (non-all-pass) filtering will shape a spectrum away from flat (white), we are well aware that if we band-pass filter white noise, we expect mild sinusoidal components at a frequency approximately that of the band-pass center. If we do this to two different white noise signals, similar weak sinusoids will artificially be created in both. Some degree of correlation will thus artificially be created. That is essentially what is going on here.

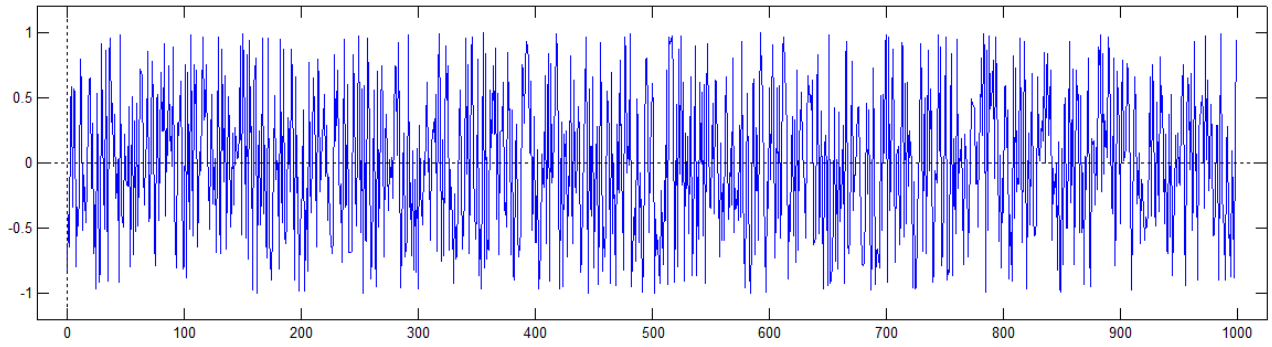
What follows in Fig. 2 through Fig. 5 should be looked at as a presentation of two different white random signals (top and bottom), generated by Matlab's ***rand*** function, as they are originally obtained (Fig. 2), along with three filtered versions. In general, staring with random noise, we often (usually?) are led to suppose something is there it isn't [3]. Famously people are not very good at producing (as in writing down numbers) random sequences or judging them as being random or not. So the best thing to do is to plot up some sequences, and recognize that probably you are not going to be happy wherever you begin and end. You keep thinking that it would be a good random sequence if only such and such did not look different from what you generally think you should expect. That is, you keep thinking that your most current example could be better. Well, it's the way it is. You just grab one (or two here) and stop worrying.

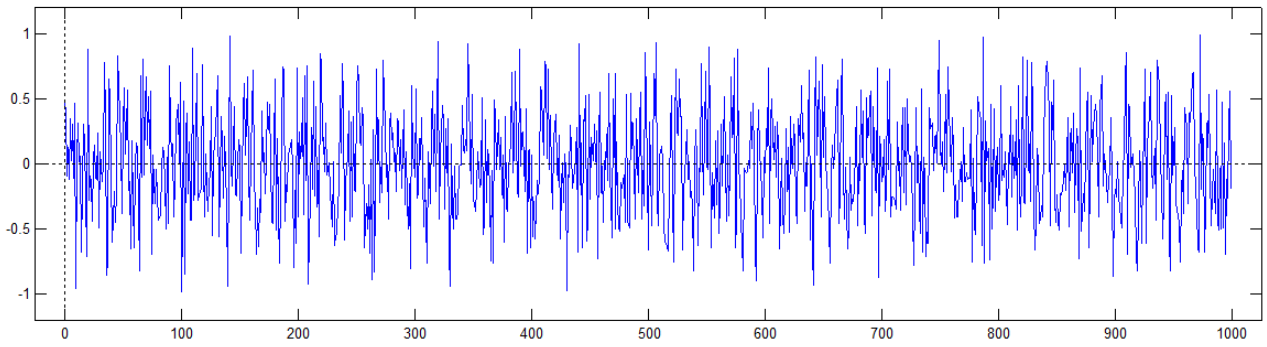
Fig. 3 shows the same two white sequences but this time after having been filtered by the "detrending" response at the top of Fig. 1. This is a high-pass response from the point of view that it nulls out DC. Otherwise is just carves some notches into the otherwise flat response (a comb filter). We do see that the waveforms are just noticeably different from those in Fig. 2, but similar to each other still. This is not really the effect of removing DC as there is essentially no DC to remove. Note that we could have tried the simple length two impulse response of (1/2)[1 -1]. The point here is that very little happens that is evident. In combination with the smoother that follows, we can see something more dramatic.

Fig. 4 is much more familiar – just a moving average which is often useful, easy to understand, and a textbook introduction to digital filtering. It is a low-pass filter, although we often want to (and can fairly easily) design much better ones. Here we are averaging 10 successive samples, and then shifting the samples by one, and so on. In this case, there is no zero at DC and we are enhancing low frequencies. We very visibly notice that the raggedness of the high frequency components is gone. Yet less that is very much periodic and existing long-term is apparent.

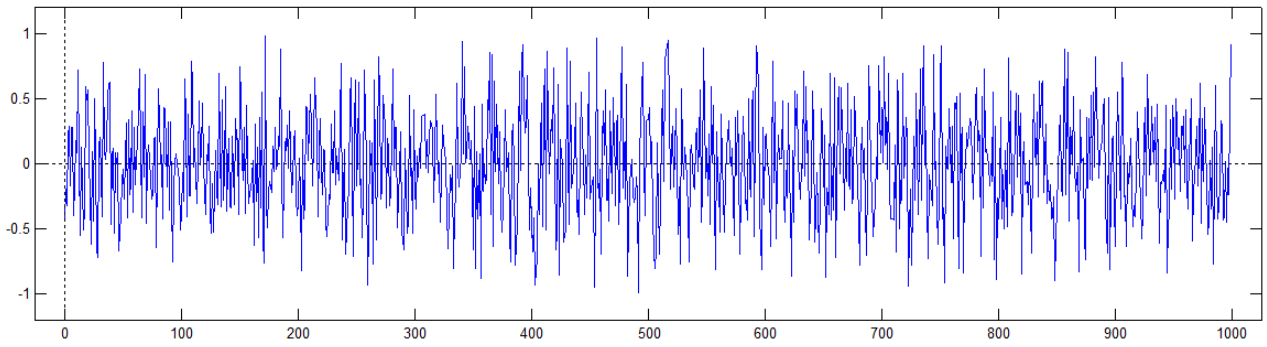


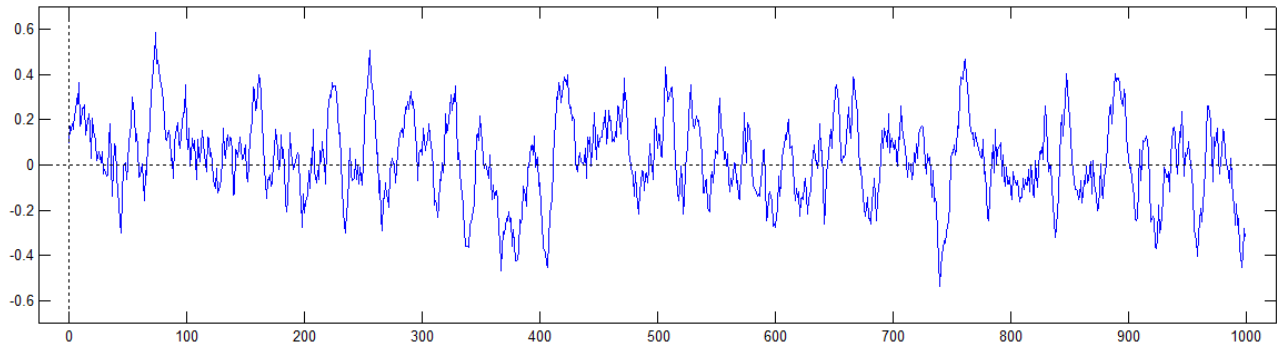
Two Original White Signals Fig. 2



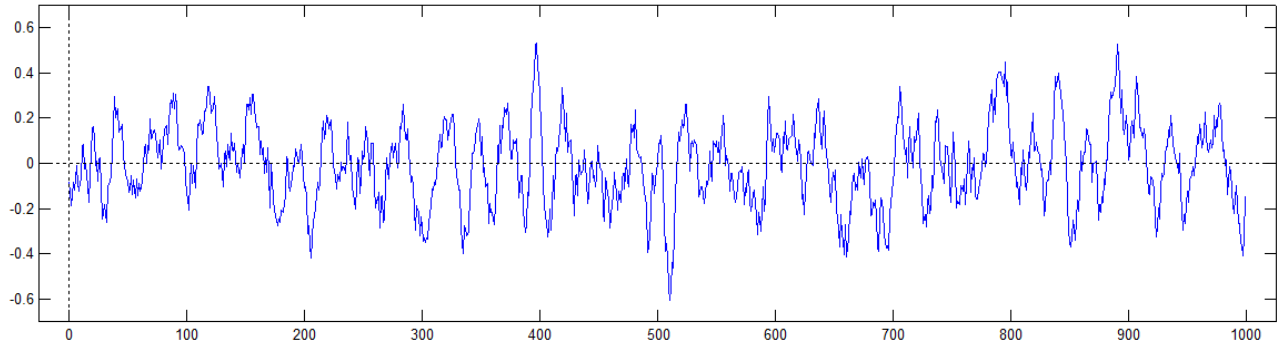


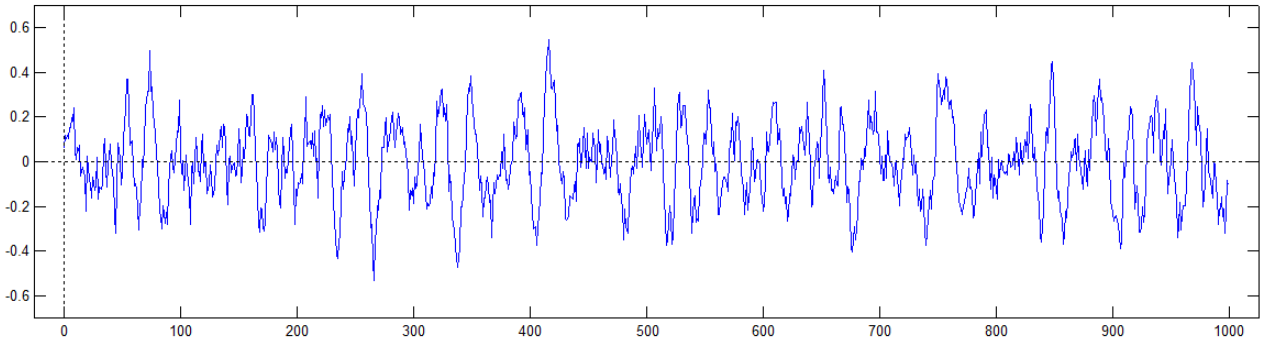
Detrended (High-Passed) Fig. 3





Smoothed (Low-Passed) Fig. 4





Detrended/Smoothed (Band-Passed) Fig. 5

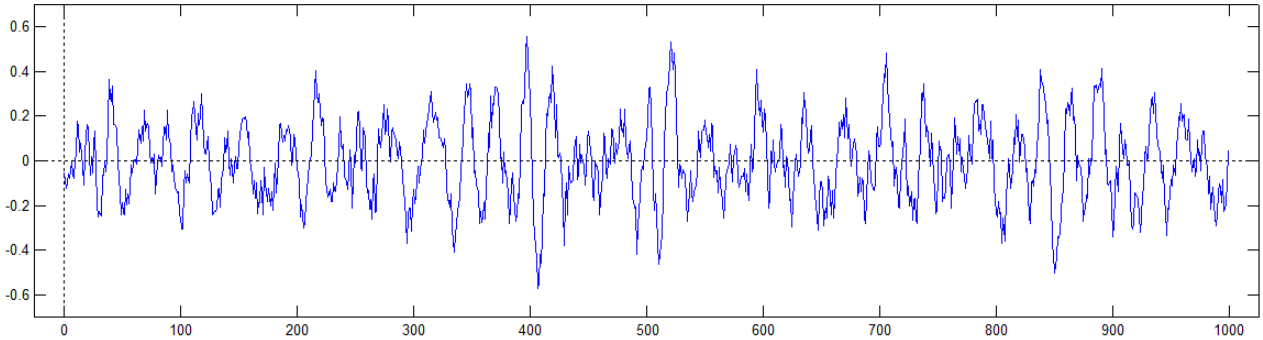


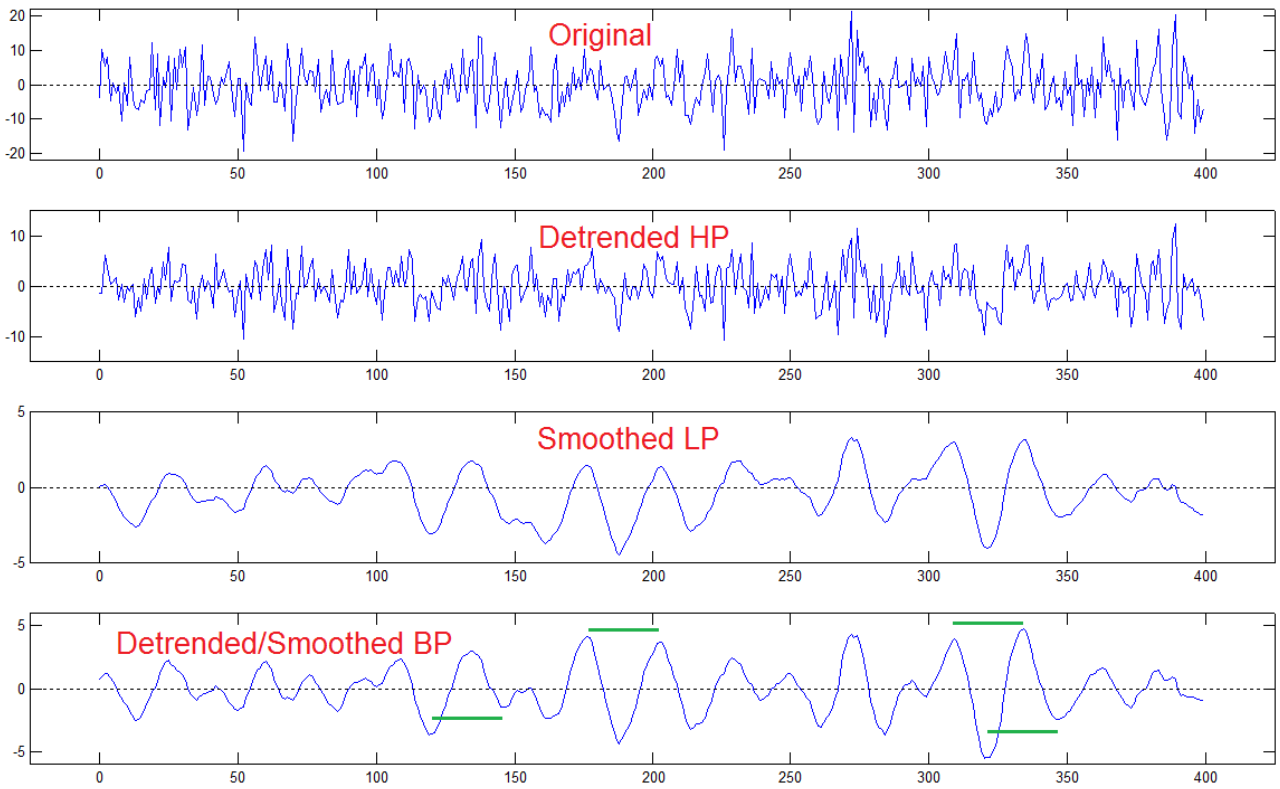
Fig. 5 shows the combination of the detrending and the smoothing. This is a case where the two signals in Fig. 5 are noticeably different from those of Fig. 4. For one thing, some low frequencies in Fig. 4 seem to go away. For example, the bottom curve in Fig. 4 seems a bit negative from samples 200-600 and then a bit positive from samples 700 to the end. The same curve, now high-passed, is seen quite straight trending in Fig. 5. The other thing is that here seem to be a lot more "cycles" trying to show in Fig. 5.

Looking back to the bottom panel (band-pass) of Fig. 1, the strongest peak is not particularly sharp but clearly located around 0.04. Since the sampling frequency is 1 (assumed normalized to 1), the frequency of 0.04 corresponds to period 25. Indeed, we do not need to stretch the truth to suggest that we see some evidence of period 25 in the waveforms of Fig. 5, at least roughly considered. Certainly these were not prominent in the white random data of Fig. 2. So we see that the two operations, each of which did shape the spectrum to other than white, combine here produce here spurious periodicity.

So far, we have specified that two different signals, as obtained by a different run of the same process (Matlab's ***rand*** function) have taken on spurious components, which are

Correlations

Fig. 6



clearly just artifacts of trying to make the data more presentable. We should be very aware that this can happen. Beyond this study however, we could perhaps have a case where two original data sets, very noisy, may or may not have not only some possible periodicities, but some correlation. [In the case of the first reference below, this was solar activity and river flow.] In seeking correlation, it is imperative that we know that the correlation is in the original data and not an artifact of some processing (however well-intended) we have done. This is hard enough without corrupting your data!

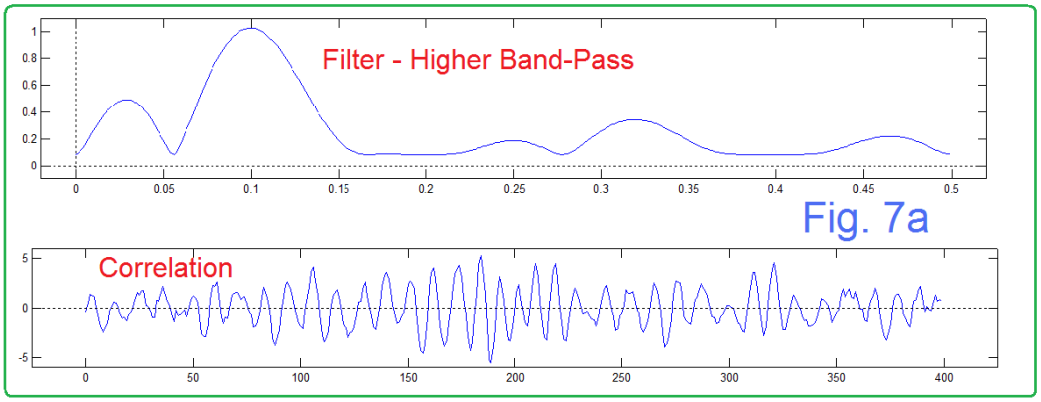


Fig. 7a

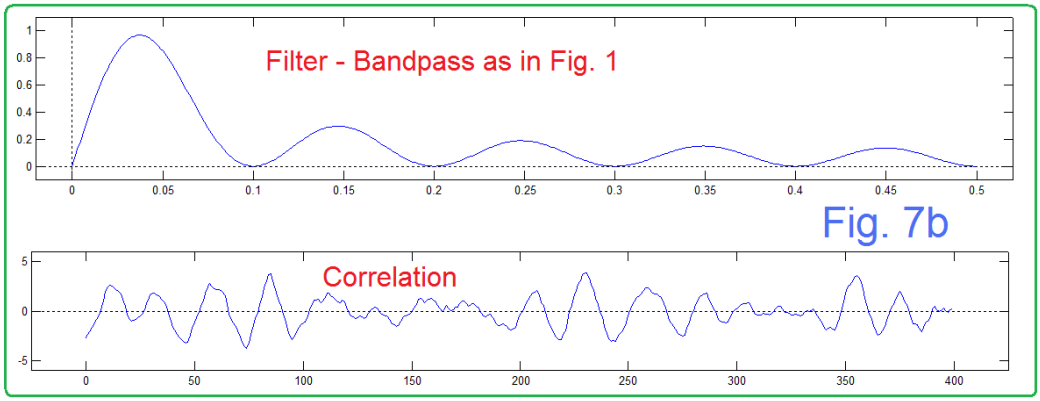


Fig. 7b

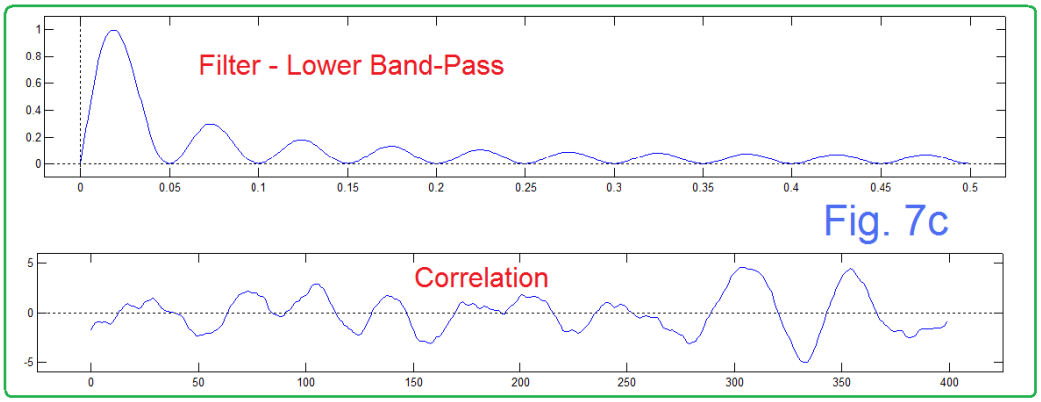


Fig. 7c

So we want to correlate the two signals, each of the pairs in Fig. 2 through Fig. 5. Well, not quite. Each of these is length 1000. Further we want to use Matlab's convolution function (***conv***) to do the correlation, so one would need to be turned around (easy enough to do) although that makes no real difference to our results. If we did convolve a length 1000 sequence with another length 1000 we get a length 1999 sequence with a generally triangular envelope as more and more samples come inside the convolution. What we want rather is for one of the sequences to be much shorter than the other. Then you do your usual convolution and keep only results where there is 100% overlap. Here we have taken 401 samples of the second sequence from 600 back down to 200, so the convolution is length 1400. Then for an output display, we keep the convolution from 400 to 799. These are numbered 0 to 399 in Fig. 6. There is nothing special about any of these choices. We did four convolutions total of the samples sequences in pairs, both of which received the same side-by-side filtering. So we were taking samples from 200 to 600 of the second sequence, and asking of there was anything similar in the first sequence. Just a basic test of the idea.

Fig. 6 shows that spurious correlation occurs at least when we get to the case of the two filtering. We see no evidence of any correlation (other than of course randomly expected) in the top panel of Fig. 6, nor is much going on apparently in the second panel. By the time we get to the third panel, the smoothed case, we immediately notice the reduction of the high frequencies. The frequencies within the first lobe of the low-pass seem to emerge. In the bottom panel of Fig. 6, where the filtering is now effectively band-pass, we see a lot more structure. Further, a lot of it seems to be close to period 25 (see four green lines of length 25, or which more could have been added. Because nothing like this is seen in the top panel of Fig. 6, we see it as an artifact of the efforts to make the presentation neater. BUT IT IS NOT REAL.

Fig. 7 is included to make the point that the peaks in the correlation really are related to the band-pass filter peak. Fig. 7b is a different example of the original case (Fig. 1, Fig. 6 bottom panel). In Fig. 7a, we have raised the band-pass; the impulse response is h = (1/12) [1 1 1 1 -1 -1 -1 -1 1 1 1 1 -1 -1 -1 -1]. The band-pass peak is at about 0.1 and the spacing of the correlation peaks is about 10. Fig. 7c lowers the band-pass peak to about 0.02, and the correlation peaks are at about a spacing of 50. The impulse response here h=(1/39)[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1] . The new filters were not derived from any particular combination of detrending/smoothing, but were just made up for the desired frequency responses. [The band-pass filtered white noise is essentially the same thing that can be used (actually to carry a melody) in a music synthesizer.] The expected result is found.

REFERENCE

[1] see “Sunny Spots Along the Parana River” as posted on WUWT on January 25, 2014 by Willis Eschenbach. Willis didn’t remember what it was called either. A commenter named Leonard Lane did. Slutsky was a Russian who published this about 1927. Yuel, better known for “Yuel/Walker” found it independently. Here is the link:

http://wattsupwiththat.com/2014/01/25/sunny-spots-along-the-parana-river/

[2] B. Hutchins , “Yearly Moving Averages as FIR Filters”, ELECTRONOTES APP NOTE 401 Dec 22, 2013
http://electronotes.netfirms.com/AN401.pdf

[3] B. Hutchins, “False Ideas About Random Sequences”, ELECTRONOTES APP NOTE 377 Feb. 2012
http://electronotes.netfirms.com/AN377.pdf

CODE

The Matlab code below made the figures of this note, with the same numbering.
Code is included here to handle ambiguity and/or provide clarification. The code for Fig. 2 was a hand modification of the h3 line. The code is not elegant !

```
% sy1

% detrender
h1=(1/2)*[1 0 0 0 0 0 0 0 0 0 -1]
% smoother
h2=(1/10)*[1 1 1 1 1 1 1 1 1 1]
% both
h3=(1/15)*[1 1 1 1 1 1 1 1 1 1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1];
H1=abs(freqz(h1,1,500));
H2=abs(freqz(h2,1,500));
H3=abs(freqz(h3,1,500));
```

```
% Plot FIlters
figure(1)
%
subplot(311)
plot([0:0.001:.499],H1)
hold on
plot([-.1 .6],[0 0],'k:')
plot([0 0],[-.1 1.2],'k:')
hold off
title('Detrender')
axis([-.02 .52 -.1 1.1])
%
subplot(312)
plot([0:0.001:.499],H2)
hold on
plot([-.1 .6],[0 0],'k:')
plot([0 0],[-.1 1.2],'k:')
hold off
title('Smoother')
axis([-.02 .52 -.1 1.1])
%
subplot(313)
plot([0:0.001:.499],H3)
hold on
plot([-.1 .6],[0 0],'k:')
plot([0 0],[-.1 1.2],'k:')
hold off
title('Detrender/Smoother')
axis([-.02 .52 -.1 1.1])
%
figure(1)
%
%
s1=2*(rand(1,1000)-.5);
s2=2*(rand(1,1000)-.5);
%
%
s1h1=filter(h1,1,s1);
s1h2=filter(h2,1,s1);
s1h3=filter(h3,1,s1);
```

```
%
s2h1=filter(h1,1,s2);
s2h2=filter(h2,1,s2);
s2h3=filter(h3,1,s2);
%
%
%
% Plot Original Random Signals
figure(2)
subplot(211)
plot([0:999],s1)
hold on
plot([-100,1100],[0 0],'k:')
plot([0 0],[-1.2 1.2],'k:')
hold off
axis([-25 1025 -1.2 1.2])
%
subplot(212)
plot([0:999],s2)
hold on
plot([-100,1100],[0 0],'k:')
plot([0 0],[-1.2 1.2],'k:')
hold off
axis([-25 1025 -1.2 1.2])
title('Original')
%
%
% Plot filtered by detrender
figure(3)
subplot(211)
plot([0:999],s1h1)
hold on
plot([-100,1100],[0 0],'k:')
plot([0 0],[-1.2 1.2],'k:')
hold off
axis([-25 1025 -1.2 1.2])
%
```

```
subplot(212)
plot([0:999],s2h1)
hold on
plot([-100,1100],[0 0],'k:')
plot([0 0],[-1.2 1.2],'k:')
hold off
axis([-25 1025 -1.2 1.2])
title('Detrender')
%
%
% Plot filtered by smoother
figure(4)
subplot(211)
plot([0:999],s1h2)
hold on
plot([-100,1100],[0 0],'k:')
plot([0 0],[-1.2 1.2],'k:')
hold off
axis([-25 1025 -0.7 0.7])
%
subplot(212)
plot([0:999],s2h2)
hold on
plot([-100,1100],[0 0],'k:')
plot([0 0],[-1.2 1.2],'k:')
hold off
axis([-25 1025 -0.7 0.7])
title('Smoother')
%
%
%  Polt filtered by both
figure(5)
subplot(211)
plot([0:999],s1h3)
hold on
plot([-100,1100],[0 0],'k:')
plot([0 0],[-1.2 1.2],'k:')
hold off
axis([-25 1025 -0.7 0.7])
%
```

```
subplot(212)
plot([0:999],s2h3)
hold on
plot([-100,1100],[0 0],'k:')
plot([0 0],[-1.2 1.2],'k:')
hold off
axis([-25 1025 -0.7 0.7])
title('Detrender/Smoother')
%
%
c=conv(s1,s2(600:-1:200));
c1=conv(s1h1,s2h1(600:-1:200));
c2=conv(s1h2,s2h2(600:-1:200));
c3=conv(s1h3,s2h3(600:-1:200));
%
%
figure(6)
subplot(411)
plot([0:399],c(400:799))
hold on
plot([-25 455],[0 0],'k:')
hold off
axis([-25 425 -22 22])
title('Original')
%
subplot(412)
plot([0:399],c1(400:799))
hold on
plot([-25 455],[0 0],'k:')
hold off
axis([-25 425 -15 15])
title('Detrender')
%
subplot(413)
plot([0:399],c2(400:799))
hold on
plot([-25 455],[0 0],'k:')
hold off
axis([-25 425 -5 5])
title('Smoother')
```

```
%
subplot(414)
plot([0:399],c3(400:799))
hold on
plot([-25 455],[0 0],'k:')
hold off
axis([-25 425 -6 6])
title('Detrender/Smoother')
```