John Menke
22500 Old Hundred Rd
Barnesville, MD 20838
301-407-2224
john@menkescientific.com

Detecting Occultations Buried in Noise

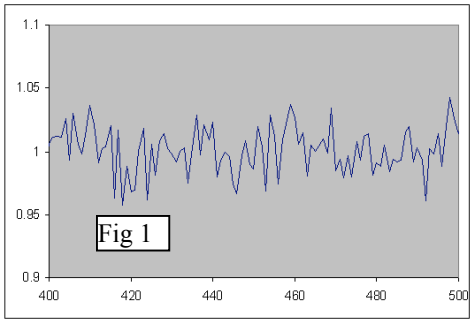
Introduction

The problem is an old one of detecting a signal buried in noise, or being sure that what you think is a signal is really there. Although this problem has received a huge amount of attention and has many elegant mathematical solutions, most of us continue to use "eyeball" estimates of probability for the detection. This works fine when there is a tidy, visible difference between the signal and the noise. However, when the signal:noise ratio is poor, we are left adrift.

Rather than solve the problem analytically, it might be useful to do some simple simulations that would help guide us, giving us a feel for the problem and its possible solutions. In this paper, I describe using Excel to generate a pseudo-data-set containing substantial noise, superimpose a signal, and then use both eye and statistical tests to determine whether a signal has been found. I then apply the same technique to investigate real world data.

Simulation

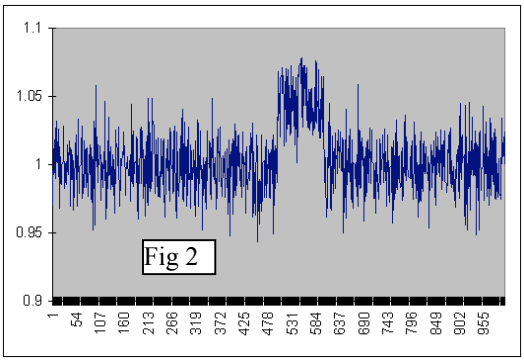
The data to be used is one that has a very similar random walk characteristic of typical video and photometric signals (details are described in the appendix). Fig. 1 shows a subset of one generated data set. The randomness is provided by the Excel Random function. The particular frequency and other characteristics of the data can be set in the spreadsheet; however, the results are not very critically dependent on the assumptions.

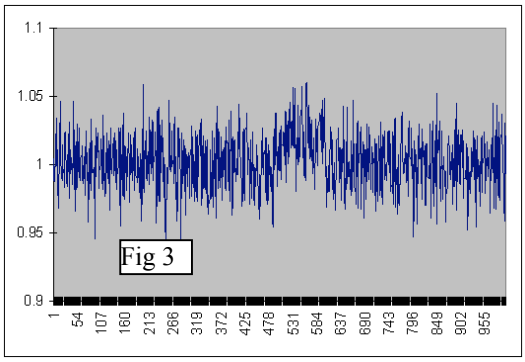


In practice, I generate a 1000-point data set, and then impose a square signal between cells 500-600 (the sixth decile). I then compare averages and standard deviations of sets of 100 cells with and without the signal cell set. Each time the recalculate button is pressed in Excel, a new set of random numbers and a new data set is generated. One can thus quickly explore whether high or low "visibility" of a signal is chance or not.

Fig. 2 shows a typical data set with large signal. The average value of the data is 1.000 and the standard deviation of the data set is 0.018. In this case, the signal level is 0.050, and the signal is obvious to see. Fig. 3 shows the same data set with a smaller signal of 0.02, roughly equal to the standard deviation in the data. With repeated excel runs, the

0.02 signal is sometimes more, and sometimes less, obvious. The issue is whether if you see a signal like that of Fig. 3, or of even less clarity, is it real?

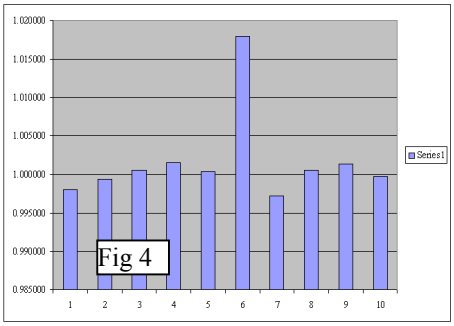




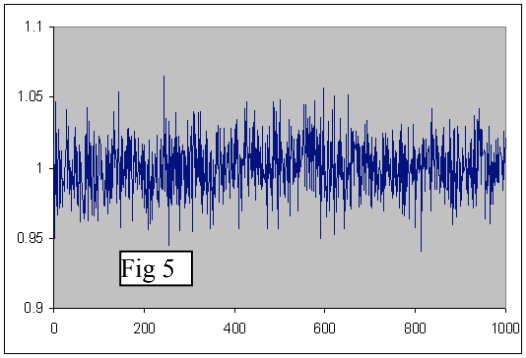
To help answer those questions, I bring in several simple measures of the data (which a user can easily generate from a similar real-world data set by using simple Excel functions). I introduce the following:

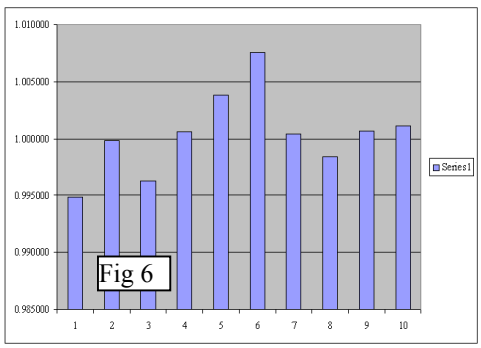
- The average value of the data for each decile
- The standard deviation of the set of ten deciles
- The standard deviation of the set of ten deciles minus the sixth decile containing the signal

For the case of signal=0.02, Fig 4 shows the averages of the deciles. Obviously, the signal stands out (even if we did not know in which decile the signal was hiding). That is because the use of the decile average lets us see the average value with an "uncertainty" of only the standard deviation of the mean, which is less than the standard deviation of the data by a factor of approximately square root of the number of points, i.e., 10x smaller. In the real world, we approach this method by using increasingly wide averages of the data to reduce the apparent effect of noise, paying for it with reduced timing precision.



In this case, repeated runs of Excel shows the bars bouncing up and down, but the sixth decile always stands out. However, with an even smaller signal=0.01, the signal is no longer obvious even on the averaged data: can one tell whether it is still there?





Well, I did cheat on Fig 6: To make my point, I chose a bar chart in which the sixth decile did not stand out. In reality, some Excel runs produce bar charts in which the signal in decile 6 stands out, while in others, like the one shown, the signal is not at all obvious in standing out from the other deciles. Technically, one can still assert that the signal is present (or absent) in the sixth decile. However, if the question is slightly different -- i.e., how sure ARE you that it is present -- then the answer has got to be "not so sure".

We can get a handle on this by another step. We can run the bar chart as shown in Fig. 6, and measure the standard deviation in the bar heights, with and without the data in decile 6. We can then run Excel ten times, which will obtain varying bar heights as the statistics bounce around. But we can then average the standard deviations to see whether there is clarity in the presence of the signal, and we can calculate the Std Dev of the various standard deviations themselves. If we do that, for a signal of 0.01, we get

Ave Std. Dev. All Deciles	.0021(5)
Ave Std. Dev. Less Decile#6	.0016(5)

There is a very statistically significant difference in the average standard deviation we get with and without decile 6. That is, the presence of the signal does quite significantly change the standard deviation of all the deciles (.0021 vs. .0016. Thus, even "buried" in 1000 points of data, a 100-point signal of 0.01 will stand out, using this measure.

How much is it standing out? The same calculations show that there is a .0005 std dev in the values of the std deviations of the deciles. That is, even if we have only one data set, rather than the ten sets (runs) used to generate these numbers, if we see a decile standing out as much as this one we would still be assured that the observation is likely positive.

Application to a Real Data Set

Fig. 7 shows a data set resulting from a video of a possible occultation by asteroid 130 Elektra on Dec. 30, 2009. The occultation was predicted to be 0.3 magnitude (30%) reduction lasting a maximum of 13.3 sec (depending on whether the observer ended up on the centerline). There was, of course, substantial uncertainty in the time of the occultation, and in the duration at my location. Observing conditions were difficult due to high clouds and bright moonlight leading to a poor signal:noise ratio. Two stars were measurable in the video, the brighter target star/asteroid, and a fainter field star. The portion of the video data analyzed lasted for approximately 60 sec. and contained approximately 2050 points, each representing one frame (0.017 sec).

The data in Fig. 7 show

Target Ave	83(60)
Field Star Ave	57(58)

Thus, the standard deviations were of the same order as the average intensity.

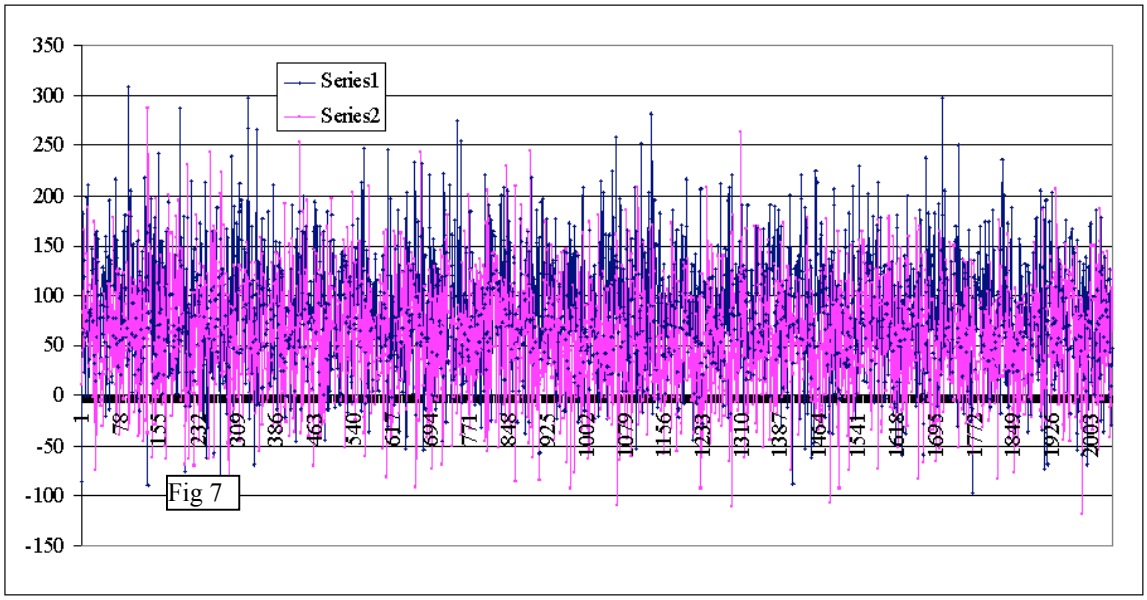
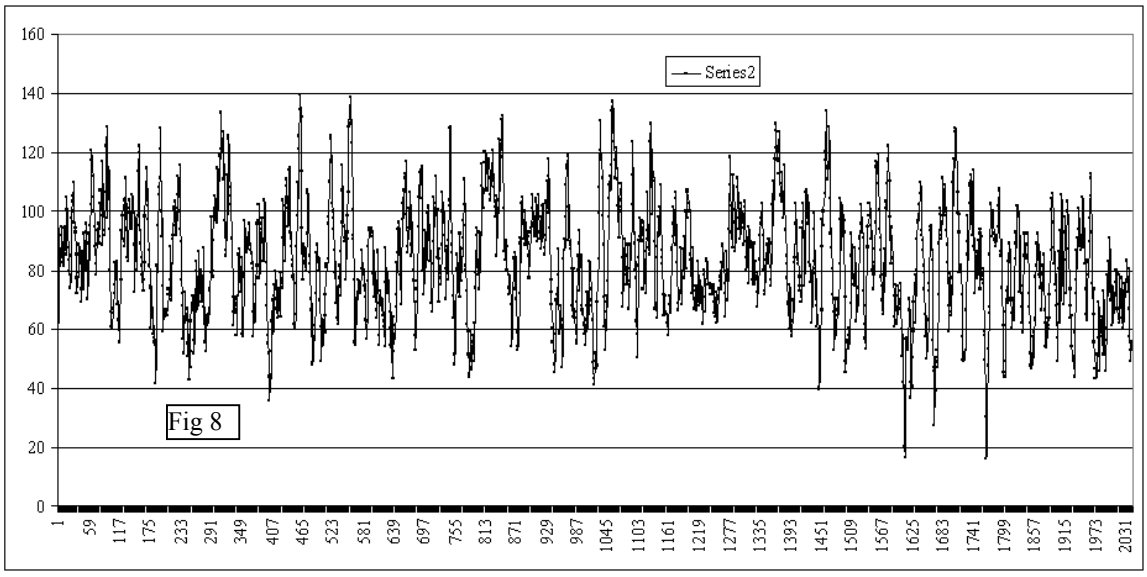


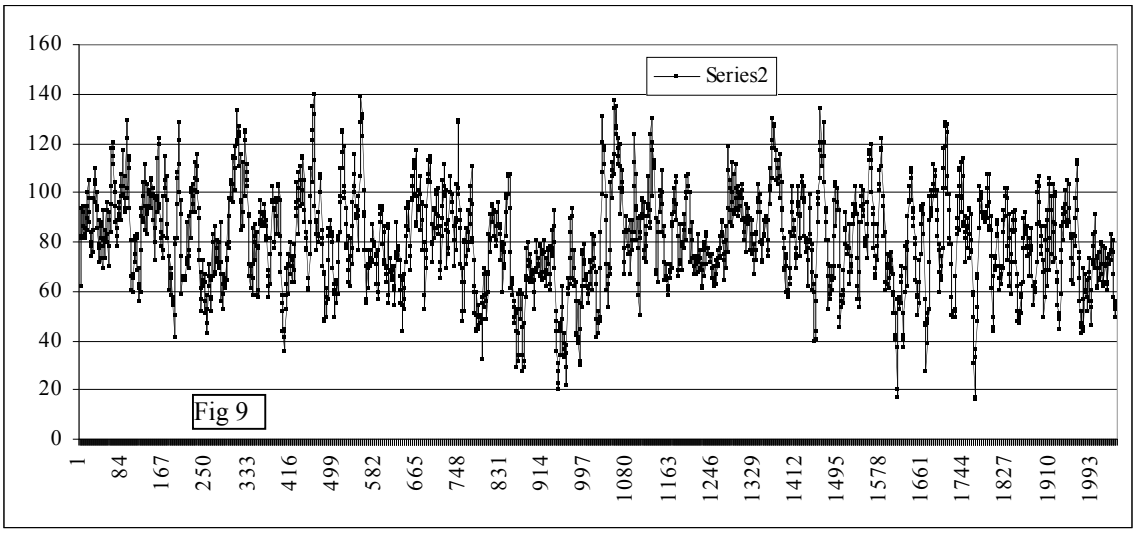
Fig. 8 shows the same data smoothed by replacing each data point by the average of the nine (9) surrounding points. The resulting data are

Target Ave	83(19)
Field Star Ave (not shown)	57(19)



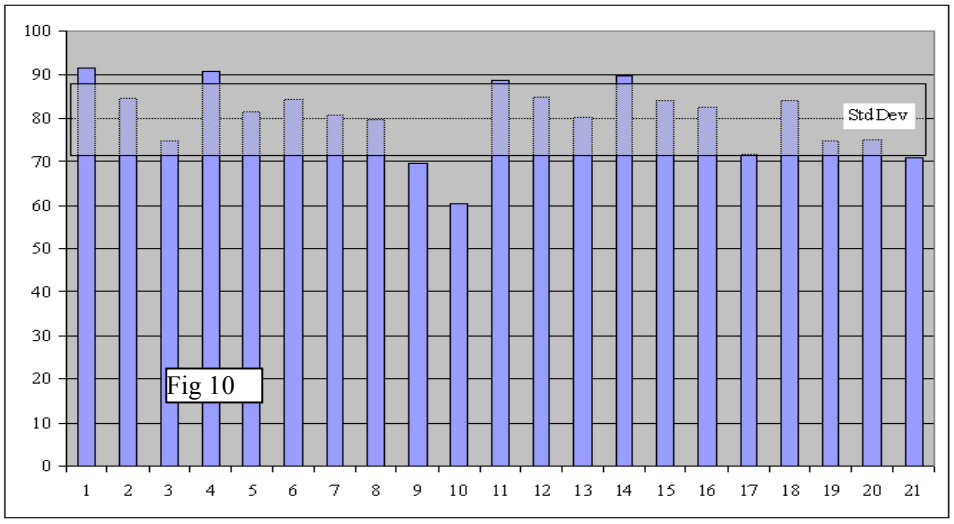
So, the question is whether there is a 30% dip lasting up to 13 sec buried in the data. Certainly, such a signal is not obviously present. We can test the visibility by adding such a signal in the spreadsheet. I used a 6.6 second signal (200 points) of 30% reduction, as shown in Fig 9. Certainly, one can see the signal present if one compares the two graphs. However, if one only saw Fig 8, one would be hard pressed (even with

this degree of averaging) to be sure the signal is present. Certainly, an even smaller signal would be even more difficult to assure. Adding to the problem is that one does not necessarily know either the time or the duration of the signal. Clearly, testing for the presence of the signal will require setting some limits on the amplitude, timing, and duration.



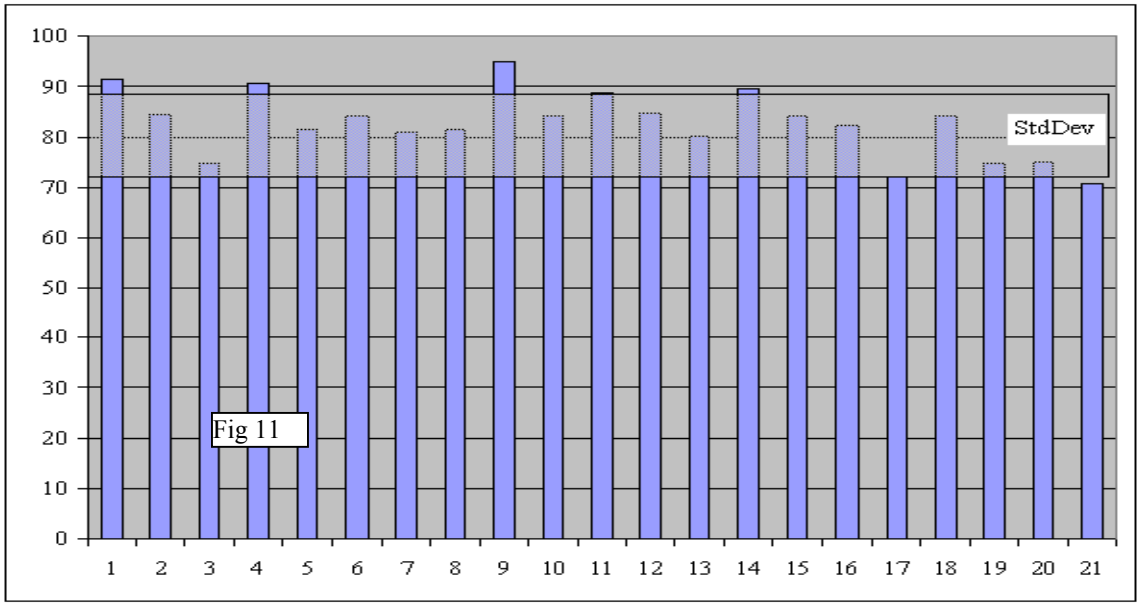
Assume that we are looking to prove whether we can see a 30% reduction (or more) lasting at least 6 sec, somewhere in the middle third of our data set. These are reasonable limits, given that we do have information, albeit with uncertainties, on each of these parameters.

Using our method of averaging selected spans of points, we can make a bar chart using 3-second (100 point) averages of the original data with an included artificial signal, to get the chart in Fig 10. The artificial signal added was 25 counts (30%), and is present in decile#9&10. The standard deviation of the bars with NO signal is +/-8 counts, and is shown by the semitransparent bar.



Setting aside deciles 9, 10, note that deciles 3, 17, 21 all are close to one std dev below the average. That is, if one were looking for one or more events of 30% lasting 3 seconds, these would be candidate events; however, they would not be very convincing. On the other hand, deciles 9,10 are both adjacent (likely from the same event spanning two bars) and both more than a std dev from the average. Thus, as a candidate for a 6 second event, one would say this would be a positive event with a good probability. Obviously, a longer duration event would be even more certain of detection.

So, was there an actual event in our data? With no artificial event in the data, the bar graph looks like Fig. 11. The only candidate area for a signal is deciles 19-20. However, this time is well outside the predicted time of the occultation, so is very unlikely. Beyond that, one can conclude that there is no occultation of 6 or more seconds having an amplitude of at least 30%. Indeed, one would conclude that even somewhat smaller events have been ruled out in the data.



Conclusion

One can use simple Excel modeling to investigate the how noise masks a signal. One can also use a simple spreadsheet approach to investigate whether a signal of a specified character is present in "real world" data. In this case, I showed that the data clearly show that no predicted signal is present.

Appendix

Here are a few hints for the Excel work described herein.

The calculation of the pseudo data stream is done by

6.746192	1.746192	1.034924
5.732455	0.732455	1.014649
5.418925	0.418925	1.008378

R1C1 is calculated as a sum of ten terms R1C1-RAND()+RAND()…
R1C2 is calculated = R1C1-5 yielding a value that bounces around zero in a Normal distribution
R1C3 is calculated = (R1C2+50)/50 yielding a value bouncing around 1.
The columns can be extended as far as desired. One can then easily add in a "signal" at any desired row.

Use of standard functions such as average and standard deviation are straightforward.

When analyzing real data, it may be desirable to average different spans of data points. Because the data may be in a long column with perhaps 2000 rows, this can be very tedious. One easy method is to set up a simple part of the spreadsheet as follows

6	E6	91.59901
106	E106	84.6396
206	E206	74.76436

The first column is a simple list using edit/fill that delineates the spans of data value entries to be averaged (e.g., from row 6 to row 106). Obviously, different intervals can be used in the column, if desired. The original data are in column E in this example. The second column is the cell reference for the data, which is put together using the concatenate function. Thus, R1C2=concatenate("E" & R1C1). The third cell R1C3=average(indirect(R1C2):indirect(R2C2)). Thus the calculated cell contents in col. 2 are called using the Indirect function from two successive vertical cells in R1 and R2. The average function then averages from cell E6 through cell E106, i.e., a span of 100 points, and puts the answer into cell R1C3. These formulae can be copy and pasted down the column to generate all the averages desired.