What is noise?

John A. Scales

Roel Snieder

Dept. of Geophysics, Utrecht University, The Netherlands

Lest men suspect your tale untrue,
Keep probability in view
— John Gay

The concept of "noise" plays a crucial role in the statistical analysis of data. As an example of a noisy record consider Figure 1 which shows the ground motion of the seismological station NE51 in St. Petersburg after an earthquake in Egypt. (In earthquake seismology periods may be orders of magnitude larger than in exploration seismology, but the principles are the same.) This time series shows no distinct arrivals or other apparent signatures of an organized nature. Given the proximity of the recording station to a major population center and to the coast such a noisy record does not seem to be very surprising.

But what is noise exactly? In the context of seismic prospecting Dobrin and Savit (1988) define noise as "spurious seismic signals from ground motion not associated with reflections." They have in mind such things as surface waves, near-surface reverberations and so on; coherent but uninteresting signal in other words. Fair enough. One might dispute the use of the term noise here, but these authors are certainly within their rights to identify certain signal as being uninteresting. But they go on to speak of "*incoherent noise*, sometimes referred to as *random noise* ... usually associated with scattering from near-surface irregularities." (Emphasis in the original.) By identifying random noise with incoherency they have sailed into rough waters. For although the signal associated with scattering from near-surface irregularities may well be incoherent (though that is debatable), it is clearly reproducible, so does it make sense to call it random? And further, there is no law that says that random processes must be uncorrelated. (Just take an uncorrelated ("white") process and apply a smoothing operator to it.)

It turns out to be extraordinarily difficult to give a precise mathematical definition of randomness, so we won't try. (A brief perusal of randomness in Volume 2 of Knuth's great *The Art of Computer Programming* is edifying and frustrating in equal measures.) In any case it is undoubtedly more satisfying to think in terms of observations of physical experiments. Here is Parzen's (1960) definition, which is as good as any:

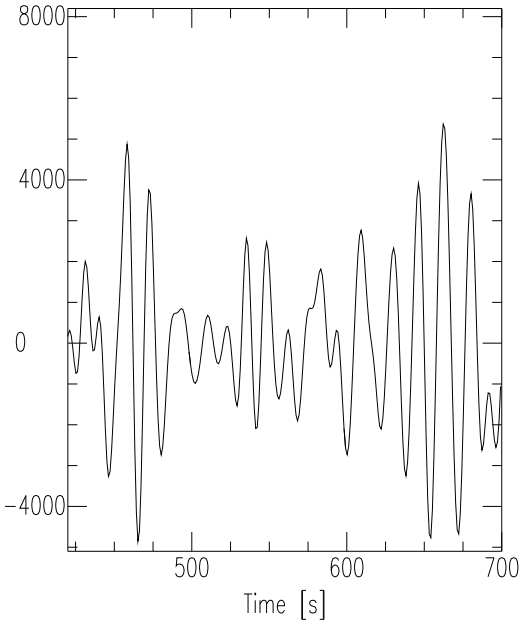

Figure 1. Vertical ground motion recorded by the NARS station NE51 in St. Petersburg after an earthquake in Egypt (22 November 1995). The time series is low-pass filtered with a corner frequency of 0.1 s.

A random (or chance) phenomenon is an empirical phenomenon characterized by the property that its observation under a given set of circumstances does not always lead to the same observed outcomes (so that there is no deterministic regularity) but rather to different outcomes in such a way that there is statistical regularity. By this is meant that numbers exist between 0 and 1 that represent the relative frequency with which the different possible outcomes may be observed in a series of observations of independent occurrences of the phenomenon. ... A random event is one whose relative frequency of occurrence, in a very long sequence of observations

of randomly selected situations in which the event may occur, approaches a stable limit value as the number of observations is increased to infinity; the limit value of the relative frequency is called the probability of the random event

It is precisely this lack of deterministic reproducibility that allows us to reduce random noise by averaging over many repetitions of the experiment. Using this definition, the "incoherent noise" of Dobrin and Savit is not random.*

But why should we care about the definition of noise? As geophysicists, the data at our disposal will always contain some features which we will not bother to explain. If we accepted our data as being absolutely precise and reproducible, then no model whose response disagreed with the observations even to the slightest degree could be correct. But we don't believe that our data are exact and exactly reproducible. And further, because we cannot calculate the exact response of our Earth models (because we cannot afford to put all the physics on the computer) and because we have only approximate models anyway (we cannot use an infinite number of parameters), there are likely to be deterministic aspects of the data which we cannot or do not want to explain. Keep in mind, however, that with enough degrees of freedom one can fit any data, even if it's not worth fitting. And the resulting model might be excessivly complicated or physically unreasonable.

In fact, in many situations "noise" is highly reproducible between different experiments and corresponds therefore to a deterministic process. For example, let us return to the seismogram of Figure 1. In Figure 2 the same seismogram is shown (on the same scale) but now the signal before the first arriving P-wave around 400 s is shown as well. The signal before the P-wave consists purely of ambient noise. It can be seen that this noise level is negligible compared to the later parts of the signal. This means that the signal shown in Figure 1 is Earth response that corresponds to a multitude of different arrivals rather than random noise. Some of these arrivals can be explained by a simple 1D earth model, but as shown by Neele and Snieder (1991) this part of the signal also contains exotic arrivals such as body waves that have been converted to surface waves. It may be that one chooses not to explain this part of the signal, discarding it as noise.

* The term "coherency spectrum" was coined by Wiener to denote the absolute value of cross covariance of two signals divided by the square root of the product of the respective autocovariance functions. Cf. Priestley (1981), page 661. If two stationary processes are uncorrelated, for example if they are independent, then the coherency spectrum is zero at all frequencies.

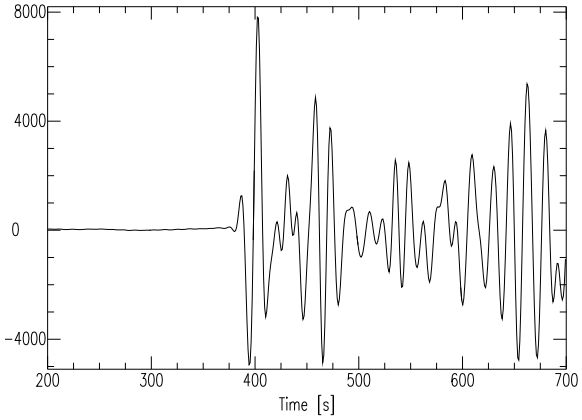

Figure 2. Same series as in Figure 1 but for a larger time interval

In doing so one must have some way of separating those data we want to explain from those we do not. So a more general definition of "noise" in the vein of Dobrin and Savit would be: noise is that part of the data which we choose not to explain.† Now, this noise may or may not be deterministic. Clearly there is unwanted deterministic data (neglected physics for instance, such as ground roll–although we personally find ground roll rather interesting, e.g. Gabriels et al. (1987)).

It is not clear to what extent deterministic signals such as unmodeled Earth response can be treated as noise. However some features in geophysical data can usefully be modeled by random processes. By this we mean the following. Suppose we isolate some part of the data that we wish to model stochastically. And suppose further that we can construct a probability law with samples that are statistically indistinguishable from the noise we are modeling. Then we can usefully call this signal random.

For instance, suppose we use a random number generator on a computer to approximate samples of a normal distribution. The algorithms used by such generators are entirely deterministic, the output being more properly described as *pseudo-random numbers*. (See Knuth (1981) for details.) Let us suppose that this random number generator is very good and its samples pass any test of Gaussianity that is known. Then, even though these numbers are clearly deterministic, they can reasonably be modeled as being Gaussian random numbers. Ulti-

† We are purposely using the term "explain" rather than "fit" since the latter seems to be burdened with certain psychological baggage. But note that fitting is nothing more than a quantitative attempt at explanation.

mately it doesn't really matter whether nature admits truly random processes or not, unless you're doing quantum mechanics.

As a more concrete example, suppose we have a well log which we have modeled as a Markovian process. In other words we have estimated the n-dimensional joint distribution function of a Markovian process, one realization of which we take to be the well log. Operationally we could do this by making histograms to approximate the 1 and 2-dimensional marginals, which are sufficient to describe a Markovian process. Then we generate pseudo-random samples of this model Markovian process. This works exactly like the pseudo-random number generator described above, but the samples come out in accordance with this Markovian distribution that we have estimated. We pass the original well log and one of the pseudo-randomly simulated logs to a fancy statistical hypothesis test and it says that at the 99% confidence level, for instance, the two are drawn from the same distribution. Then, whether or not we believe the real log is a realization of a random process, we can usefully model it as such.

So what does all this have to do with geophysics? There are three main implications:

- **Stacking of data**: The notion that averaging over repeated realizations of an experiment (stacking) reduces noise (compared to signal) presupposes that the noise in the different experiments is uncorrelated because only then do we get the desired noise-suppression. The criterion here is that the correlation in the noise is zero for different experiments.
- **Prescription of a-priori errors in Bayesian inversion:** In such an inversion we need to prescribe the joint distribution function of the data errors (e.g., the data covariance matrix and mean if the distribution is assumed to be Gaussian). It is fine to include signal-generated "noise" in this (although a purist might argue against this). However, this type of noise will in general be highly correlated (between different samples, between different shots, and between different receivers). The correlations are crucial here, but in practice they may be difficult to quantify (Gouveia and Scales (1997)).
- **Making the decision how well to fit the data:** In a least-squares fitting procedure, one must use use chi-square or some other measure of the misfit to determine how well the model explains the data. Here, one needs to know which parts of the data are real (or interesting), and which part should be considered noise. (The latter should not be fitted.) This is not only an issue of noise levels or correlations (e.g., knowing the mean and variance of Gaussian noise), but in some applications it is important to identify data and noise in a more subtle way. For example, ground roll may be considered noise that should not be fitted by weird reflectors.

This means that the issue "what is noise" is of more than academic interest. If we define noise as being that data we choose not to fit, then we must have a model that explains the rest of the data. If not, this could be a sign that the "noise" is carrying important information. The association of noise with non-deterministic processes may be misleading since the concept of noise is also used as the garbage-can of unexplained deterministic phenomena. A treatment of this type of noise on purely statistical grounds may lead to conceptual as well as practical problems.

References

Dobrin, M. and C.H. Savit, *Introduction to Geophysical Prospecting*, McGraw Hill, 1988.

Gouveia, W. and J.A. Scales, Bayesian seismic waveform inversion: parameter estimation and uncertainty analysis, *JGR, 103,* no. B2, 2759-2779, 1998.

Gabriels, P., R. Snieder and G. Nolet, In situ measurements of shear-wave velocity in sediments with higher-mode Rayleigh waves, *Geophys. Prosp., 35*, 187-196, 1987.

Knuth, D., *The Art of Computer Programming*, Vol II, Addison Wesley, 1981.

Neele, F. and R. Snieder, Are long-period body wave coda caused by lateral heterogeneity?, *Geophys. J. Int., 107*, 131-153, 1991.

Priestley, M.B., *Spectral Analysis of Time Series*, Academic Press, 1981.

Parzen E., *Modern Probability Theory and its Applications*, Wiley, 1960.

Note:

This paper is to appear in GEOPHYSICS, July/August 1998. It is one of a series of three articles, by Scales and Snieder.