Retrospective correlation

Francesco Mulargia
Dipartimento di Fisica, Settore di Geofisica, Università di Bologna, Italy

Abstract
One of the main tools in phenomenological studies is the identification of correlations among different processes. This is essentially effected in retrospective with the specific aim of finding a positive result, and that leads to a parameter optimization which introduces a bias, so far only marginally considered, in the significance level of the results. If the correlation can be validated in a forward study in which parameters are kept fixed, such a bias is irrelevant. Unfortunately, forward studies are often infeasible for either cost or intrinsic reasons. This is the case of geophysics, due to the comparatively long time scale of recurrence of the phenomena. Unbiased estimates can be obtained in retrospective if each of the optimal choices is properly identified and accounted for. An estimate of the bias is made in a specific case, which can be written in closed form. While simulation confirms its good performance, the latter shows that apparently highly significant retrospective correlations may be insignificant.

Key words *earthquake prediction – statistical analysis*

1. Introduction

Testing the validity of a hypothesized correlation deals primarily with establishing if it occurs more often than chance. The first problem is then how to define «chance». In geophysics, this is an often encountered problem, and a typical case is that of earthquake precursors. In the latter case, to gauge the correlation against «chance» the set of the hypothesized precursors is typically tested for performance against both the real catalog and a randomized seismic catalog. The problem is complicated due to the fact that randomized catalogs cannot be uniquely defined since the present knowledge of earthquake physics does not allow either a deterministic or a statistical satisfactory model to be derived (Kagan, 1994; Ben Menahem, 1995).

Mailing address: Prof. Francesco Mulargia, Dipartimento di Fisica, Settore di Geofisica, Università di Bologna, Viale C. Berti Pichat 8, 40127 Bologna, Italy; e-mail: mulargia@ibogfs.df.unibo.it

However, the lack of unique models is not the major cause of difficulties in establishing phenomenological correlations. There is a problem of greater importance, which has so far received only marginal consideration. This is the bias in the correlation estimates. In fact, the common procedure is to analyze data in retrospective and optimizally select (rather than keeping *a priori* fixed or randomly choose) all free parameters in order to achieve the best apparent result, *i.e.* the correlation the least apparently probable by chance. The problem can be readily solved if the data are originated by repeatable experiments, because retrospective analysis can be used to cast an issue of correlation which can then be validated in an independent, typically forward, study (Mulargia and Gasperini, 1996). On the contrary, if only retrospective studies are feasible, a substantial distortion of reality in favour of the correlation can be expected. This is the case of some geophysical phenomena such as earthquakes and volcanic eruptions, since the comparatively long time scale of geologic processes would require waiting at least several decades for the forward occurrence of a sufficient number of events.

Unbiased retrospective estimates are possible if all the steps in parameter optimization are identified and properly taken into account in the number of favourable *vs.* number of total cases. This task is a difficult one because the tailoring of some parameters is often implicit, such as in the selection of the region in which the correlation is operative. In geophysics this selection is justified in terms of «tectonics», a term that implies more or less subjective *ad hoc* choices, which are not explicitly discussed (*cf.* Keilis-Borok *et al.*, 1988).

The present paper is devoted to estimating the importance of the bias induced by retrospective adjustment of parameters. This will be achieved by studying correlation in a particular simple case, regarding a single variable, as in the case in which two series are analyzed by considering only the times of occurrence of the events. In such a case correlation is commonly defined as *association*.

2. Retrospective association

In studying the association between two series of events, the typical record consists of a number m of type a events, and of a number n of type b events, in a time interval of length T time units (fig. 1). The series are said to be associated in time on the basis of the number of events b following the a events by a fixed lag u, *i.e.* when b events occur within windows of amplitude $2h$ centered u instants after an event a. For simplicity, let us assume that the number of correlated events s is fixed *a priori*. This number is ideally equal to the minimum between m and n, but nobody would dismiss a correlation occurring on, say, $s = 35$ among two series of $m = 43$ and $n = 52$ events. With this constraint, the only variable that can be optimally chosen is then the lag u, while the probability of finding s events associated in a window of width $2h$ depends on the specific model assumed. To proceed further let us therefore first consider this.

2.1. *Assumption of a specific model*

A simple, yet widely applicable, model (Cox, 1955) puts no constraint on the events a, assumes that events b follow a simple (stationary) Poisson process, that $m2h$ is small with respect to the total length T of the period analyzed, and that $2h$ is small enough that the oc-

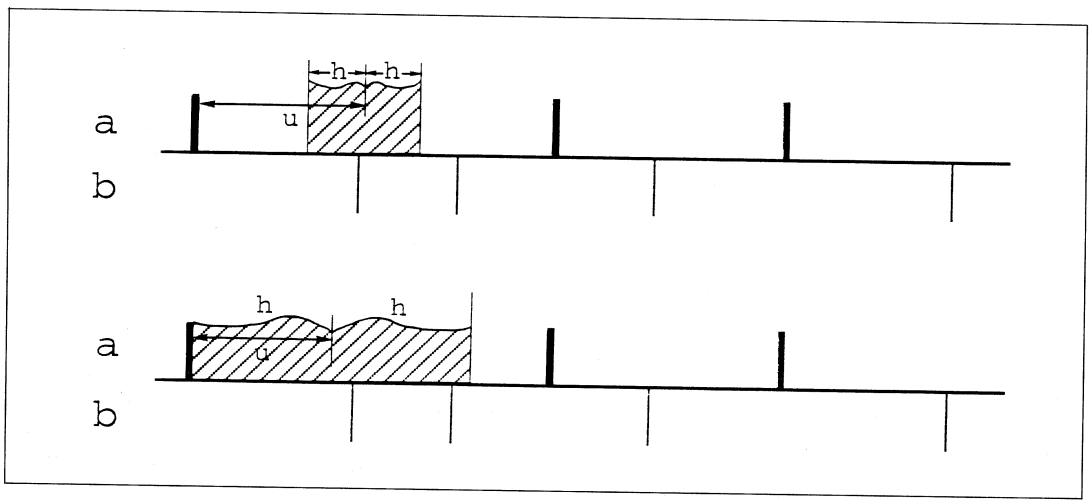

Fig. 1. The association parameters between a series of events a and a series of events b are the lag u and the window width $2h$. The general case is shown at the top, the case $u = h$ is shown at the bottom.

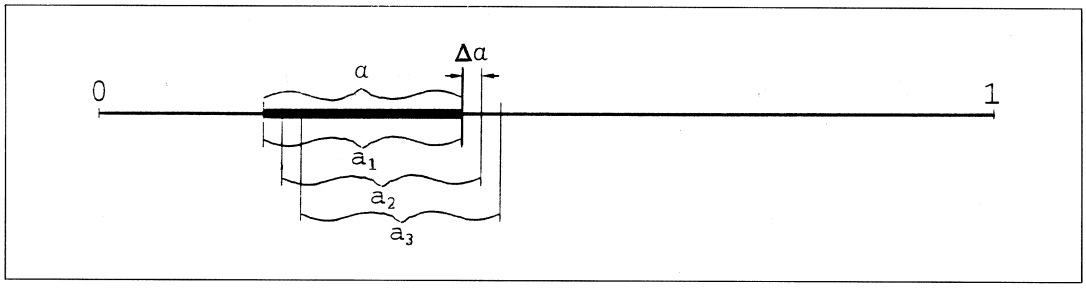

Fig. 2. Partition of the unit probability interval with equivalent association intervals.

currence of two a or two b events within $2h$ can be neglected. Then, the probability of occurrence of one b event within $(u-h,\ u+h)$ from an a event is approximately equal to $m2h/T$, and the probability that x events a and b are associated with parameters u, h is binomial

$$p(x) = \binom{n}{x} (m\,2h/T)^x\,(1 - m\,2h/T)^{m-x}. \quad (2.1)$$

Note that if a lack of constraints were desirable on the b series, rather than on the a series, the reasoning could be reversed by simply using negative u values.

3. Bias introduced by the optimal choice of u

We have fixed s, and found the corresponding best h from the data once given the distribution. But we have not fixed the lag u, which is indeed the very target of any correlation analysis. This is chosen by retrospective optimal selection on a specific realization, *i.e.* the available data set. Should we have another realization, we would choose another optimal value of u. This optimal selection introduces a bias in the correlation estimates, which, according to eq. (2.1), are valid if u was kept fixed *a priori* at each realization. In other words, eq. (2.1) was valid if u had been determined through an independent set of measurements.

Let us now estimate the retrospective bias. Let us then keep h fixed and assume that a particular realization has its best association with a number s of a and b events associated in the window $(\tilde{u}-\tilde{h},\ \tilde{u}+\tilde{h})$, corresponding to an apparent low probability of random occurrence equal to $\tilde{p}$. Actually, we would have been equally content with finding the same low probability of association with any other different value of u. To evaluate the importance of this effect it is convenient to transfer the reasoning on the unit probability interval (0,1).

Consider first the case in which the number of associations s is equal to min $\{m, n\}$, and let the corresponding probability be $\tilde{p}$. We can then draw on the unit probability interval (0,1) the set α corresponding to the probability that one b event falls in the window $(\tilde{u}-\tilde{h},\ \tilde{u}+\tilde{h})$ which is $(\tilde{p})^{1/s}$ (fig. 2). A set of other equivalent intervals, all with the same probability of containing one b event, and thus with the same $\tilde{p}$ for s events, can then be drawn by shifting the set α by fixed steps of amplitude $\Delta\alpha$. A total of $\tilde{j} = 1 + (1-\alpha)/\Delta\alpha$ are needed to cover the whole probability interval (0,1). Obviously, the coverage will not be perfect, but the remainder can be controlled by choosing small values of $\Delta\alpha$, like $\Delta\alpha = (1-\alpha)/1000$. Now consider the fact that one is not interested in the single event A_1; {s events in the interval a_1 of length α}, but in the union of the events $A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_{\tilde{j}}$, *i.e.* in finding s events in any of the equivalent intervals $a_1, a_2, \ldots, a_j$

each one of length α. The probability of the union of the A_i events is

$$P = S_1 - S_2 + S_3 - ..... + (-1)^{(i+1)} S_i + ... S_{\tilde{j}} \tag{3.1}$$

where

$$S_1 = p\{A_1\} + p\{A_2\} + ... + p\{A_{\tilde{j}}\},$$

$$S_2 = p\{A_1 A_2\} + p\{A_1 A_3\} + ... + p\{A_{\tilde{j}-1} A_{\tilde{j}}\},$$

$$S_3 = p\{A_1 A_2 A_3\} + ... + p\{A_{\tilde{j}-2} A_{\tilde{j}-1} A_{\tilde{j}}\},$$

$$.....,$$

$$S_{\tilde{j}} = p\{A_1 A_2 ..... A_{\tilde{j}-1} A_{\tilde{j}}\}.$$

This expression is insensitive to discretization, provided that $\Delta\alpha$ is small (see above). The assumption $s = m$ drastically simplifies calculus for all the joint events. To see this, consider that the only way the joint event $A_k A_l A_m$ can occur is that the s events take place in the interval $a_k a_l a_m$, which is the intersection of the three intervals, and this intersection interval is the same as that of the event $A_k A_m$. The probability of the joint events $A_k A_l A_m$ and $A_k A_m$ event are thus the same, but they enter P with opposite signs since they belong, respectively, to S_3 and S_2, and they cancel out.

The resulting expression for P is then simply

$$P = \sum_{1}^{j-1} \alpha^s - \sum_{1}^{j-2} (\alpha - \delta\alpha)^s \simeq \alpha^s \left[1 + s\frac{\delta\alpha}{\alpha}(j-2)\right]. \tag{3.2}$$

The correctness of this result was verified in extensive simulation.

If $s \neq m$ no such simplification occurs and calculus becomes very heavy, making simulation preferable. However, also simulation quickly becomes costly if the number of events in each series is larger than 10. Fortunately, when the latter occurs, the probability that the intersection of the events $A_i A_k ..... A_r$ may occur outside the intersection of their intervals of definition $a_i a_k ..... a_r$ becomes very small, and the above equation provides again a good approximation. Simulations with $s \neq m$ gave analogously good matches, provided that $m \geq 10$.

4. The practical relevance of retrospective bias: an example

Consider for example the case of two completely random (*i.e.*, simple Poisson) series of 12 events each, *i.e.*, $m = n = 6$ and let us take $s = 7$. What is the probability that inspecting them *a posteriori* by searching for the best correlation a significance level α is found? The problem was tackled in simulation, with 100000 realizations. The results are shown in table I, where it is easy to see that the significance level calculated according to the unbiased estimate eq. (2.1) is always much smaller than its corresponding random definition, *i.e.* the probability of encountering by chance such a result. Obviously, the value calculated according to eq. (2.1) would be correct if the value of u had been fixed *a priori* and kept unchanged rather than optimally selected at each realization. The bias, *i.e.*, the discrepancy between the value of the significance level (which is by definition the probability of encountering such a result by chance) and the observed relative frequency (which is the empiri-

Table I. The observed relative frequency of significance levels equal to a given value («experimental» values) calculated according to eq. (2.1) («uncorrected» results) does not match, as it should by definition, the value of the significance level. This is due to nonrandom retrospective optimal sampling, which introduces a bias. The latter leads to a blow-up factor which, calculated according to eq. (3.2), and applied as a correction to the results of eq. (2.1), yields the «corrected» results.

Uncorrected	Experimental	Corrected
0.001	0.02	0.016
0.005	0.06	0.05
0.01	0.11	0.08
0.05	0.29	0.24
0.10	0.42	0.38

cal definition of probability), is the effect of the retrospective optimal choice of u. The blow-up factor calculated according to eq. (3.2) is also shown. It is immediately seen how the theoretical correction agrees well with the observed frequency, and provides a correct estimate of the blow-up factor. Note how the true significance levels are an order of magnitude or more larger than the uncorrected levels even with such small values of s, m, and n.

5. Discussion and conclusions

The standard practice in correlation analysis is to retrospectively choose all free parameters in order to obtain the best apparent association. This optimal choice introduces a bias in the estimates. If the association can be validated in a forward study in which parameters are kept fixed, this bias is irrelevant. Unfortunately, forward studies are often infeasible for their cost, or intrinsically impossible for the long time scale of evolution of the phenomena, as mostly happens in astronomy and geophysics. In this case, unbiased estimates of the significance of the association may be obtained by identifying and accounting for the optimal choices implicit in the retrospective analysis. The explicit study of one particular case shows that the blow-up factor can be written in a simple closed form. It also shows how this factor is very large and, unless corrected, may easily lead to conclude as highly significant correlations which are, indeed, insignificant.

Acknowledgements

I am indebted to Professor Michele Caputo for many stimulating discussions. This work was performed with contributions CNR Gruppo Nazionale Difesa dai Terremoti and MPI 60% and 40%.

REFERENCES

BEN-MENAHEM, A. (1995): A concise history of mainstream seismology: origins, legacy, and perspectives, *Bull. Seism. Soc. Am.*, **85**, 1202-1205.

COX, D.R. (1955): Some statistical methods connected with series of events, *J. R. Stat. Soc. B*, **17**, 129-164.

KAGAN, Y.Y. (1994): Observational evidence for earthquakes as a nonlinear dynamic process, *Physica D*, **77**, 160-192.

KEILIS-BOROK, V.I., L. KNOPOFF, I.M. ROTWAIN and C.R. ALLEN (1988): Intermediate-term prediction of occurrence times of strong earthquakes, *Nature*, **335**, 690-694.

MULARGIA, F. and P. GASPERINI (1996): Precursor candidacy and validation: the VAN case so far, *Geophys. Res. Lett.*, **23**, 1323-1326.