Geophys. J. Int. (2001) **146,** 489–496

Retrospective selection bias (or the benefit of hindsight)

Francesco Mulargia

Settore di Geofisica, Dipartimento di Fisica, viale Berti Pichat 8, 40127 *Bologna, Italy. E-mail: mulargia@ibogfs.df.unibo.it*

Accepted 2001 March 23. Received 2001 March 20; in original form 1999 July 15

SUMMARY

The complexity of geophysical systems makes modelling them a formidable task, and in many cases research studies are still in the phenomenological stage. In earthquake physics, long timescales and the lack of any natural laboratory restrict research to retrospective analysis of data. Such 'fishing expedition' approaches lead to optimal selection of data, albeit not always consciously. This introduces significant biases, which are capable of falsely representing simple statistical fluctuations as significant anomalies requiring fundamental explanations. This paper identifies three different strategies for discriminating real issues from artefacts generated retrospectively. The first attempts to identify *ab initio* each optimal choice and account for it. Unfortunately, a satisfactory solution can only be achieved in particular cases. The second strategy acknowledges this difficulty as well as the unavoidable existence of bias, and classifies all 'anomalous' observations as artefacts unless their retrospective probability of occurrence is exceedingly low (for instance, beyond six standard deviations). However, such a strategy is also likely to reject some scientifically important anomalies. The third strategy relies on two separate steps with learning and validation performed on effectively independent sets of data. This approach appears to be preferable in the case of small samples, such as are frequently encountered in geophysics, but the requirement for forward validation implies long waiting times before credible conclusions can be reached. A practical application to pattern recognition, which is the prototype of retrospective 'fishing expeditions', is presented, illustrating that valid conclusions are hard to find.

Key words: earthquake physics, selection bias.

INTRODUCTION

The natural phenomena occurring in the solid Earth are rarely simple in nature. In general, they are the sum of so many different non-linear effects that writing their constitutive equations is a formidable task. Phenomenological studies therefore play a fundamental role in their investigation. This has been the traditional approach of geology, later inherited by geophysics, in that it applies the instruments and analytical tools of physics to the same questions. As in all phenomenological studies, success is possible if a wealth of accurate data are available. This is the case in much exploration geophysics and in global seismological traveltime studies, which have enabled accurate mapping of the velocity structure of the interior of the Earth. However, this success also exposes some of the pitfalls of all issues rooted in data. Statistics has developed very effective methods of sampling integrated with data analysis, which are the bases of modern agricultural sciences, pharmacology, production control, etc. (*cf.* Cochran & Cox 1957). Such techniques, which may be categorized under the heading of *experimental design*, assume the ability to acquire new data at will and to control the experimental conditions. Both of these are often impossible in geophysical studies. Data availability introduces major constraints and this is one of the crucial problems in many fields of geophysics. This lack of control on the 'experiment', and the scarcity of data, open the door to *selection bias*.

In general, once identified, selection bias can be corrected, or at least its extent can be evaluated, allowing the accuracy of the analysis to be properly defined. The most dangerous type of selection bias is therefore that which is not identified. This can occur even in careful studies, and a notable example can be found in recent observational seismology. It regards seismic tomography, which appeared to have a resolution sufficient to infer the fine detail of the topography of the core–mantle boundary (Morelli & Dziewonski 1987). Unfortunately, these authors neglected selection bias due to the non-random spatial distribution of the seismic stations, which are mostly deployed on the continents (Stark & Hengartner 1993).

In some cases, research is focused on phenomena where limited understanding is combined with difficulty in acquiring new data, rendering all analysis essentially empirical and retrospective. In this case, a particular type of selection bias may

be largely concealed in the process of data analysis itself, and goes under the name of *retrospective selection bias*. Earthquake physics, of which even the basic processes are poorly known (Ben-Menahem 1995), and for which only semi-qualitative models have so far been derived (Main 1996), provides a typical example. The present article seeks to analyse retrospective selection bias by examining the statistics of earthquakes and of the onset of volcanic eruptions, which share similar underlying physical processes of fracture. We make explicit reference to earthquake and eruption prediction, since their extreme character readily exhibits the problems of retrospective analysis.

In retrospective studies, data are analysed with the specific aim of 'finding something'. The methods of analysis differ according to whether a sound theory is available or not. In the first case, a mathematical description exists, the validity of which has been previously established by extensive comparison with experiment. New data can then be used to further confirm or to extend the applicability of that theory. In the second case, where no valid theory exists and there is only some vague idea of where to look and what to look for, the analyst explores the available data to find support for a credible theory. This is not as straightforward as it seems. While the vast majority of the studies on earthquake physics follow some existing 'model', and therefore appear to be framed according to the first model, this is hardly ever true, because most such models are only semi-qualitative. There is one notable exception, the universally accepted slip-on-a-plane-fault earthquake model, which treats ambient stress as the main variable. It provides constitutive equations for the radiation field in the low-frequency limit, which have been validated by many people. Can this be taken as a valid model for the physics of earthquakes? The fact that faults display a fractal geometry, that is, they are as far as possible from planar, is strongly suggestive that it cannot. A more definite proof is provided by the unquestionable failure of attempts to predict earthquake occurrence deterministically. All of these theories are based, more or less closely, on the slip-on-a-plane-fault tenet and regard stress as the main determinant.

In conclusion, there is as yet no consolidated model to be confirmed in earthquake physics, and all current studies are in the domain of attempting to find a model, with a vague idea of what to look for. This 'fishing expedition' approach will stop only when a viable theory is found. The researcher and the fisherman care only about the final result, and the successful reporting of a scientific explanation erases most traces of the process of data selection which the researcher has used to achieve his apparently positive result. This is pernicious because all the options examined and discarded as unproductive are not counted as part of the statistical analysis, introducing a crucial bias in the calculated probability that the supportive evidence is due to chance.

RETROSPECTIVE BIAS

Deriving a novel result from data implies the observation of evidence that is 'anomalous' with respect to a general situation of 'normality'. 'Normality' is described in terms of a number of variables, the values of which will follow an assumed distribution. The observation of measured values in the tails of such a distribution is at the basis of identification of any 'anomaly'. Note how these anomalies are 'outliers' to the distribution, the rejection of which is carefully controlled since each outlier is a potential 'discovery'. The probability of any measured value lying in the tail of a distribution can be estimated if (i) the distribution is known and (ii) the sample is random. A wealth of data is usually available for the 'normal' situation, so that the first requirement (a known distribution) should be easily verified. The major difficulty lies in the second requirement, since the retrospective researcher is *not* dealing with a random sample. His analysis consists of a careful scrutiny of the data expressly aimed at identifying subsets that do not exhibit 'normal' behaviour, that is, he considers sets that are just the opposite of a random sample. Estimating probabilities of occurrence of deviations from 'normality' on the basis of non-random sampling gives rise to a severe *retrospective selection bias* in favour of the apparent anomaly. For example, if a large number of cases, say 1×10^5, are drawn from a truly random population, events with a very low probability of occurrence of, say, 1×10^{-4} can be expected to occur about 10 times by mere chance. Let us examine this case in more detail.

To a first approximation, the intervals of definition of the different parameters can be assumed as mutually independent, so the total number of possible choices is given by the Cartesian product of the number of possible choices for each parameter. Formally, denoting by r_i the ith parameter (out of a total of N), which defines the working set, and assuming that its interval of definition has n_i values (typically $n_i = 2$, i.e. a lower and an upper bound) and that each of the latter is chosen among m_{ij} values, the total number of cases considered, N_T, is

$$N_T = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \prod_{k=1}^{m_{ij}} r_{ijk} \,. \tag{1}$$

This means that the retrospective researcher may inadvertently consider a very large number of cases, and this, together with the disregard of having done so, leads him to conclude that he has observed 'unlikely' cases occurring 'surprisingly' far more often than expected.

For example, consider the following scenario: the occurrence of earthquakes in region A is linked to the occurrence of earthquakes in region B. This conclusion is based on retrospective examination of a seismicity catalogue after selecting a lower magnitude threshold, that is, one value chosen among the three values 3.0, 3.5 and 4.0. Each region is then specified as a polygon with four vertices, each identified by two geographical coordinates, each of which is examined for four possible values. This constitutes an unstated optimal choice amongst 32 values per region. Finally, the association is tested based on selecting events within a time interval of the catalogue, employing five possible values for the starting point of the interval and five for its end. The total number of cases considered in this simple example is, according to eq. (1), 2.4×10^3, so by mere chance we might expect to find one association with apparent probability of the order of 10^{-4}, some tens of cases with probabilities of the order of 10^{-3}, and so on.

It must be noted (i) that the parameter choices will be dependent on one another to some extent, and (ii) that retrospective selection is seldom conducted efficiently and exhaustively through a systematic exploration of the parameter space. It is common practice to rely on 'intuition' to reduce the size of the set that must be explored by trial and error to achieve the best apparent result. These constraints produce a retrospective bias that is smaller than the theoretical limit of an efficient optimal selection but by an unknown amount.

THE DIFFERENT FACETS OF STATISTICAL DATA ANALYSIS

In the following, we discuss the various aspects of retrospective selection bias, which are largely concealed by the procedure of defining the 'anomalies' themselves, and against which standard statistical analysis is ineffective (Stark & Hengarter 1993). To illustrate the traps of retrospective investigation, at the end of the discussion we present a practical example regarding an application of pattern recognition—the prototype of empirical retrospective analysis—to the eruptive activity of Mount Etna.

In the most classical treatment, statistical data analysis is ignored altogether and the data are interpreted subjectively. In this case, the reliability of the conclusions depends essentially on the intuition and on the authority of the proponent.

Nowadays, virtually all data are analysed using statistical tools. Basic techniques, such as simple regression, correlation and spectral analysis, are widely known and commonly used, but their implications and fine details are less well understood, and this often leads to misapplication, misinterpretations and incorrect results (Mulargia & Gasperini 1995; Gonzato *et al.* 1998). Classical advice on data analysis (Fisher 1935) is to concentrate on ideas and not on the method, which should be the simplest and most intelligible that does the job. When more refined statistical tools are used, more detail can be revealed. While these tend to be more powerful, and possibly to be used more carefully, they also tend to divert interest towards the technique of analysis itself. This distracts the author first, and the reader later, from the data set and its definition. Furthermore, use of these refined techniques may impress the reader with their complexity, and discourage him from investigating further. The results may then be uncritically interpreted as reliable. Estimates based on refined computer simulations are particularly exposed to this danger, unless the source codes are made available (which happens very rarely).

FIGHTING RETROSPECTIVE SELECTION BIAS: STRATEGY 1, THE 'HOPEFUL' APPROACH

We now analyse how the problem of retrospective selection bias can be defined in detail and tackled. We outline our attempt to control retrospective selection bias by describing three specific strategies.

If one could count *all* the cases explicitly and implicitly considered, it would be possible to make an unbiased estimate of the probability of occurrence of the 'anomaly'. This is a particular case of the general problem of hypothesis testing, which is discussed later in the paper. Statisticians have worked on multiple hypothesis testing for some time and call it the *false discovery rate* problem. A number of procedures are available for controlling it statistically (Hsu 1996). In short, this is a multiplicity problem associated with the simultaneous test of not one but many hypotheses. Improved Bonferroni methods have been proposed by Simes (1986) and Scheffé (1959), and, with a broader scope, by Benjamini & Hochberg (1995). Satisfactory solutions to specific problems have also been developed (Eckhardt 1984; Stark & Hengartner 1993; Mulargia 1997).

Unfortunately, many of the options considered are declared only implicitly, and are sometimes hidden in the procedure of analysis itself. Explicitly optimal selections are usually exhibited as clearly subjective and are the easiest to spot and correct. The task of identifying optimal selections becomes more difficult when these are still explicit, but less evident. In this case, a variety of nuances are possible, starting with weak arguments such as 'precursory phenomena [are] observed at very large distance, and [are] absent in the vicinity of the focal region' (Varotsos & Lazaridou 1991; Thurber & Sessions 1998). Next come geographic arguments such as 'this study reports about earthquakes in Central California' (Keilis-Borok & Rotwain 1990). In general, unless quantitative arguments are provided, for example, about the resolving power of the seismic network used, it is not clear why earthquakes should be grouped spatially using state borders.

Arguments at the edge of subjectivity are more subtle. Tectonic arguments are a good example. A study concerning earthquakes on the San Andreas fault (an apparently sound problem) will typically imply selecting events in a geographical region of complex geometry, the shape of which is chosen on the basis of arguments concerning the accurate spatial identification of active faults, a goal that is impossible to achieve objectively (e.g. Knopoff *et al.* 1996). Since each vertex fixes two parameters, and since polygonal regions with more than 10 vertices are common (Knopoff *et al.* 1996), the selection of the operative set is likely to conceal an optimal selection of a large number of parameters (see the example above).

An obvious countermeasure is the systematic use of stability studies, but this has its own problems. How should the stability analysis be performed? How large is the variation range? How large is the tolerance? Sound results must be robust against variations so large that they would most likely already have been 'discovered'. Similar arguments apply to the other parameters for selecting the size window and the time window, and the difficulty is to identify each single subjective choice (Mulargia 1997).

Identifying optimal selections hidden in the definitions is easy only in principle, since these can easily be overlooked. For example, choosing 'cumulative moment as the variable to study the features of a given seismic region and its correlation with other regions' appears to be a sensible choice for the intrinsic physical meaning of this variable. However, since moment release is dominated by large events, the author may, in effect, have reduced a massive catalogue of some tens of thousands of events to the few largest ones. The subjective operation of choosing the boundary of the chosen region may result in a revised choice of these large events, with substantial effects on the results.

Finally, identifying and correcting the optimal selections hidden in the procedure of analysis itself is possible only in some particularly simple cases such as those described above. In general this is very laborious, often requiring specialist mathematical methods (e.g. Stark 1992).

In light of the above, it is clear that correcting all the possible processes of optimal selection is a formidable task. It becomes possible if the selection procedure is exhaustive and efficient. If, as mostly happens, rather than a systematic exploration of the parameter space, selection is conducted by intuition and trial and error, it is hardly possible to track the number of choices that have really been explored. All that can be said is that this number is smaller than the theoretical limit of an efficient retrospective selection.

FIGHTING RETROSPECTIVE BIAS: STRATEGY 2, THE ULTRA-CONSERVATIVE APPROACH

Realizing that retrospective bias is ubiquitous, that it may crucially affect most analyses, and that correcting it *ab initio* is very difficult, we examine the radical alternative of rejecting all apparent anomalies except those that have an extraordinarily low retrospective level of significance. The idea is that since retrospective optimal selections induce much lower apparent significance levels, rather than accepting as 'anomalies' cases beyond, say, 2 or 3 standard deviations, one moves the acceptance threshold to 6 sigmas or so, that is, to apparent significance levels of 10^{-6} or lower (Anderson 1990; Mulargia 1997). This may appear as a wise and practical option. Unfortunately, it also has the unwanted effect of discarding possibly genuine 'anomalies'. To examine this, consider the typical case of testing a variable for which some apparent 'anomaly' has been found retrospectively.

A DRAWBACK OF STRATEGIES 1 AND 2: THROWING THE BABY OUT WITH THE BATHWATER

We discuss in the following a basic statistical issue that, in spite of being well known, is very often overlooked. Let us hypothesize that some as yet undiscovered process representing an 'anomaly' with respect to known 'normal' evidence does indeed exist, and let this be characterized by average values of some variable x. Irrespective of the form of the distribution of x, which in geophysics is often non-Gaussian, the central limit theorem (Mood *et al.* 1974) tells us that the average of the anomalous x will have a sampling distribution that is asymptotically Gaussian with parameters μ_A and standard deviation σ_A. These can be identified as 'anomalous' since the same average for the 'normal' situation will also likely follow a Gaussian distribution, but with parameters μ_N and σ_N.

Our task is to identify the 'anomaly' in the data. In general, the 'normal' population will be documented by a large number of samples n_N of s units each, so that we may assume that μ_N and σ_N are known. On the other hand, the 'anomalous' population is likely to be documented by much smaller samples, possibly just one sample—a single datum. This is the *case history* approach, quite common in geophysics (Wyss & Martirosyan 1998), which describes a single piece of evidence at a time.

Following standard procedure, we identify an 'anomaly' on the basis of a sample with average values of x equal to $\bar{X}_A$ by testing the null hypothesis $H_0 : \bar{X}_A = \mu_N$ against the (one-sided) composite alternative $H_1 : \bar{X}_A > \mu_N$. The test will be based on the z standard variate,

$$z = \frac{\bar{X}_A - \mu_N}{\sigma_N / \sqrt{n_A}}, \tag{2}$$

where n_A is the number of samples in the 'anomalous' set.

Rejecting H_0 for positive z values larger than a given threshold Z means accepting the risk of disregarding all the values larger than Z that are indeed 'normal'. This is the *type I* error, which concerns the misidentification of a true null hypothesis. That is, we identify a truly normal event as 'anomalous'. Probabilities α of type I error (α is called the *significance level* of the test) no larger than 0.01 (which in a two-sided test is equal to 3 standard deviations), or at most 0.05 (which in a two-sided test is equal to 2 standard deviations), are commonly adopted.

Significance level: the lower, the better?

If one independently knew the parameters of the 'anomalous' population μ_A and σ_A (as happens in established clinical testing), one could test the null hypothesis $H_0 : \bar{X}_A = \mu_N$ against the simple alternative $H_1 : \bar{X}_A = \mu_A$. In this case, $\bar{X}_A$ will identify (see Fig. 1) a right tail in the 'normal' distribution, corresponding to the type I error discussed above, and a left tail in the 'anomalous' distribution, corresponding to the probability β of a *type II* error, with β estimated by

$$z' = \frac{\bar{X}_A - \mu_A}{\sigma_A / \sqrt{n_A}}. \tag{3}$$

Rejecting H_1 for all values of average x smaller than $\bar{X}_A$ will mean accepting a risk of rejecting true anomalies equal to β. This is conveniently described in terms of the *power* of the test, which is defined as $1-\beta$, and should ideally be close to 1 (Lehmann 1986). Any attempt to lower α, that is, to shift $\bar{X}_A$ towards the right, will imply an increase in β. In other words, guarding against false retrospective issues may lead us to reject genuine 'anomalies'.

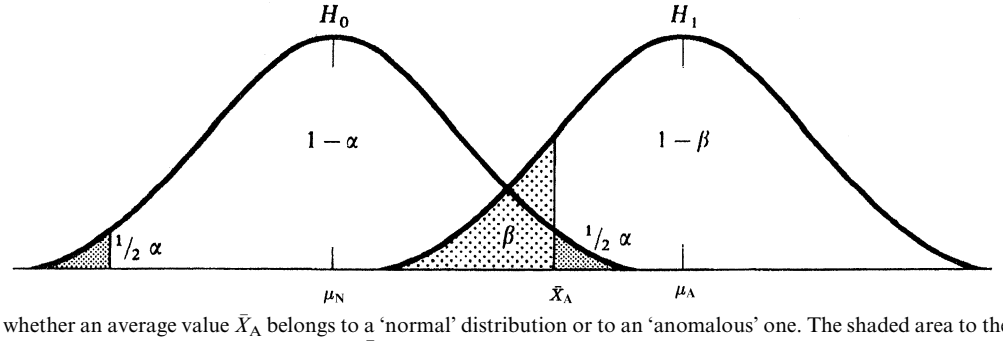

Figure 1. Testing whether an average value $\bar{X}_A$ belongs to a 'normal' distribution or to an 'anomalous' one. The shaded area to the right of $\bar{X}_A$ under the 'normal' curve represents the probability of declaring $\bar{X}_A$ to be 'normal' and is equal to the significance level. The dotted area to the left of $\bar{X}_A$ under the 'anomalous' curve represents the probability of declaring $\bar{X}_A$ to be 'anomalous' and 1 minus its value is equal to the power of the test. Larger values of $\bar{X}_A$ have a lower significance level, but also a lower power. The only way to achieve simultaneously low significance level and high power is to separate the 'normal' and 'anomalous' curves from each other by increasing the sample size (see text).

The key role of sample size

According to eq. (2), the value of $1/z$ and therefore the values of α and β vary as the inverse square root of the total number of anomalous samples collected, n_A in a set of size N. If this is large, we may be able to meet the goal of having a low significance level and, at the same time, a high power. In other words, if the number of 'anomalous' data is large, in principle it is possible to meet the requirement of Strategy 2 for retrospective studies without compromising the capability of detecting genuine anomalies. For example, if some 'precursor' were observed in retrospect for 10 000 earthquakes it could be quite comfortably accepted as a precursor, irrespective of the optimizations that led us to identify it.

The situation worsens progressively when the number of events decreases. The worst case is reached when case histories are considered. There is then no average over samples but instead a single datum. This has two pernicious effects: first, there is no help from the central limit theorem in guaranteeing that the sampling distribution tends towards the Gaussian, so any assumption of Gaussian approximation is inappropriate. A non-Gaussian, most probably non-parametric, approach can be attempted, but statistical methods do not handle single points well. Second, no control whatsoever is possible on the power of the test, and the only way to achieve low significance levels is to accept that the test has very low power. In light of this, we see that case histories have little hope of being effective analytical tools. They are by definition mere case reports, and should only serve to arouse interest in some precise topic, thereby promoting detailed systematic research. In no way can particular case histories be used to draw firm conclusions.

A practical example of the effects of 'anomalous' sample size: fumarole temperatures

In order to see the beneficial effect of sample size let us consider, as an example, the case of the 'fumarole' temperatures at a hypothetical volcano. This case considers the 20 yr record of daily readings at 10 different fumarole sites. On this basis we know that the average 'normal' temperature is 450 °C, with a standard deviation of 100 °C. During the period in question, a major eruption occurred, and a retrospective analysis of the available record apparently suggests that the average temperature of the 10 fumaroles in the week prior to the eruption was 500 °C. How should one interpret this 'anomaly'? Although temperatures can only be positive in value, their sampling distribution should nevertheless be approximately Gaussian. In this case, the sample is made of a single point, and all reasoning is speculative. Using the equations above and disregarding all retrospective bias, we have that the sample difference in the average temperatures corresponds to $z = (450-500)/(100/\sqrt{1})$, that is, an apparent retrospective (one-sided) significance level of 0.31. Obviously, there is no way to exploit this datum in any significant analysis, illustrating the point made above that case histories can only reasonably be used to promote further research.

Alternatively, let us hypothesize a situation involving a larger sample. Let us assume that we have a series of 22 eruptions, and that before them on average the same temperature of 500 °C (average over the 10 fumaroles) as above is measured. Now eq. (2) yields an apparent significance level of 0.01. If all the parameters for choosing the sample had been selected at random, this value would suggest the existence of a temperature anomaly before the eruptions. However, some retrospective optimal selection has almost certainly occurred, so that this level appears insufficient to support any anomaly issue.

Let us then say that, in agreement with Strategy 1 or 2, to counter the retrospective bias a more conservative threshold for significance level equal to 0.00001 is chosen, which corresponds to a (one-sided) z value of 4.27. This would obviously mean that the above set of 22 temperatures with average 500 °C can be comfortably judged 'normal'. However, let us assume that an 'anomaly' does indeed exist and that its population mean and standard deviation are respectively equal to 515 °C and 100 °C. By accepting this sample as 'normal' we are then likely to make a type II error with an even larger probability than a type I error. In fact, since $z = (500 - 515)/(100/\sqrt{22}) = 0.704$, the value of β is equal 0.242. We therefore face the following dilemma: the significance level must be lowered to counter retrospective bias, but this may lead us to throw away some genuine results.

Eqs (2) and (3) suggest the key to the solution of this quandary: increase the size of the 'anomalous' sample to meet simultaneously the desired levels of α and β. For example, leaving all the above temperature values unchanged, but changing the size of the 'anomalous' set to 73 units, will set the significance level below the required 0.00001 level and, at the same time, increase the power to 0.90. If increased temperatures were observed routinely in advance of 73 eruptions, we could with confidence conclude that these temperatures were anomalous. In other words, if sample size is large enough to allow an exceedingly low significance level and, at the same time, a high power, then we can comfortably reject the null hypothesis of 'normality'.

Note that using the values above we would guard against misclassification of 'anomalies' with average $T = 515$ °C, but not, for example, against those with $T = 507$ °C. Since we do not generally know the parameters of the 'anomalous' distribution, the customary approach (Scheffé 1959) is to estimate the power of the test as a function of this temperature, using the procedure we have just outlined, assuming a variety of values for the mean of the 'anomalous' distribution (e.g. 505, 508, 511, 514, 517, 520 °C, etc.) together with a fixed sample size and standard deviation. This produces a set of β values that allows us to estimate the interval in which the anomalous temperatures would be successfully detected, so that independent evidence can then be used to validate the results.

FIGHTING RETROSPECTIVE BIAS: STRATEGY 3, THE TWO-STAGE APPROACH

We have seen that a large sample size, in connection with one of the above strategies, can effectively control retrospective bias. However, large samples are rare in geophysics, which often involves small, and often very small, sample sizes. For example, the number of instances of recurrence of large earthquakes on the same fault segment is, at best, a handful (Pantosti *et al.* 1993). The eruptive episodes at a given volcano are, at best, a few tenths, etc. We ask whether retrospective bias can be tackled with the further constraint of small size. Miracles are not to be expected, but a wise approach seems to be to get the most out of the few available data.

Namely, all available retrospective data are analysed, operating the usual approach—avoiding inconsistencies and untenable *ad hoc* selections (such as extracting individual data points as mentioned at the beginning of Strategy 1)—without correcting for any optimization. Standard test and significance levels are then adopted, with no attempt to validate the data. These are used to establish 'candidate' anomalous results that are then investigated within a second separate stage of forward validation. In other words, once the rules for defining the best retrospective anomaly have been determined, rather than attempting to estimate the retrospective bias explicitly (as demanded by Strategy 1) or to protect against it by adopting very conservative significance thresholds (as demanded by Strategy 2), we use standard or marginally conservative significance levels such as 0.01 or 0.001 to identify the *candidate* 'anomalies'. These establish an unequivocal definition of the 'anomaly', that is, of the laws of the game. Keeping these laws strictly unchanged, *validation* is then performed on a separate, truly independent set. This must necessarily consist of a further experiment run forward in time (Mulargia & Gasperini 1996). The great advantage is that standard significance tests can be employed, because in forward time we do not have the benefit of hindsight, and there is no need to account for retrospective bias. As a consequence, less stringent requirements are imposed on sample size. For example, in the fumarole case examined above, a sample size of 22 would be sufficient to establish a retrospective candidacy, and another forward sample size of 22 to validate it, a total of 44, approximately half the number required to establish the anomaly by retrospective validation alone.

The limitation of this strategy is that due to the comparatively long timescales of many geophysical phenomena it may nevertheless require a comparatively long time to acquire forward data, the only data that are guaranteed not to suffer from retrospective selection bias. On the other hand, there seems to be no easy alternative: significant sample sizes are required to draw any conclusion. The relative sizes can be calculated using eqs (2) and (3).

This strategy can produce interesting issues in the case of negative results, that is, when candidate 'anomalies' cannot be identified, as discussed in detail in the following paragraph.

There is no hope without a sound candidacy

Suppose that during retrospective analysis of a given set of data in the framework of the two-stage candidacy–validation strategy a researcher is unable to find any clear sign of anomaly according to the standards for candidacy, that is, below the 0.01 (or 0.05) significance level. In this case, there can be little hope that a forward test, without the help of retrospective optimization, will support any specific thesis. The dispassionate researcher will therefore simply discard the 'anomalous' issue and redirect his efforts in other directions. Curiously, this is at odds with what generally happens. Sometimes researchers 'fall in love' with their hypotheses and defend them by re-tuning their data selection [for instance, compare Varotsos & Lazaridou (1991) with Varotsos *et al.* (1993), or Thurber (1996) with Thurber & Sessions (1998)].

As a consequence, application of the candidacy procedure, that is, identifying and correcting the most obvious optimizations and applying significance analysis to standard levels, is sufficient to undermine the credibility of many apparently established hypotheses. The characteristic earthquake and the time- and slip-predictable earthquake models, to name a few, do not pass the standard tests for candidacy (see respectively Kagan 1993 and Mulargia & Gasperini 1995).

Practical application to pattern recognition, the prototype of retrospective selection bias

The term *pattern recognition* indicates a generic search for structure in a given set of data, with little or no reliance on existing models. As such, it represents the prototype of 'fishing expeditions'. Pattern recognition at some level is used in all types of analyses. The oldest, and still most popular, method of pattern recognition is based on visual examination and subjective intuition, medical diagnosis being the typical example. As we mentioned above, this (at least in part) subjective approach explores only some of the possibilities. On the other hand, exhaustive procedures of pattern recognition have been developed thanks to modern computing technology. A wide variety of algorithms have been developed to this end (Duda & Hart 1973; Bezdek 1987; Fukunaga 1990), but applications in geophysics have so far been limited. The most well-known are probably those of the Russian approach to predicting earthquakes (Gelfand *et al.* 1978; Keilis-Borok *et al.* 1988; Kossobokov *et al.* 1990), which rely on long-established logical algorithms rather than the more recent computationally intensive procedures.

Both types of algorithms have been applied to identify the patterns of seismicity preceding and accompanying the eruptive activity of Mount Etna (Mulargia *et al.* 1991, 1992). These applications have led to the conclusion that there is a link between the occurrence of local earthquakes and the flank eruptive activity of Mount Etna, while the summit activity seems to be uncorrelated. In particular, it was found that local seismicity accompanies each eruption, with the occurrence of at least seven events within an 80-day interval centred at the onset of each eruption. The statistical pattern recognition approach used by Mulargia *et al.* (1992) estimated this pattern to be significant in retrospect below the 0.05 level for the 11 flank eruptive episodes during the period 1974–1989. The pattern correctly identified all events and misidentified as 'eruptive' another 13 intervals. The identification was also checked in retrospect for stability, and proved highly stable since truncating the learning set as far back as 1979 December 31; that is, covering only 40 per cent of the total period left the recognized pattern unchanged. The conclusion that flank eruptive activity is tied to local tectonics at Etna seemed therefore to emerge as a sound candidate.

We examine whether this issue stands against validation by posterior occurrences. During the period from 1989 December 31 up to the time of writing, only one flank eruption occurred at Mount Etna. This started on 1991 December 14 and lasted for 15 months, with a total erupted volume of 250×10^6 m^3 (Gresta, personal communication, 1999). In this same period, instrumental problems limited the available seismic record to the interval 1990 January 1–1996 December 31. The results of the validation attempt are as follows. The onset of the 1991 eruption was characterized by seven seismic events within the prescribed interval. Entering this value in eq. (2) together with the values of the 'normal' population inferred from the learning period 1974–1989, i.e. $\mu_N = 3.95$, $\sigma_N = 3.81$, yields a probability of a type I error (falsely rejecting the null hypothesis) equal to

0.21. At the same time, entering this same value in eq. (3) together with the values of the 'anomalous' population inferred from the learning period, i.e. $\mu_A = 13$, $\sigma_N = 7.63$, yields a probability of a type II error (falsely rejecting the alternative hypothesis) also equal to 0.21. In other words, the single forward occurrence appears inconclusive and neither confirms nor negates the recognized pattern. At the same time, extending the learning set to 1996 December 31, with the inclusion of the 1991 eruption, introduces minor modifications in the parameters of the 'normal' and 'anomalous' populations and does not suggest any instability or change in behaviour. Specifically, extension of the learning period shifts μ_N from 3.95 to 4.1, σ_N from 3.81 to 3.54, μ_A from 13 to 12.5 and σ_A from 7.63 to 7.48. In short, the validation attempt gives inconclusive results, and the recognized pattern remains merely a candidate. This example also illustrates once more that sound conclusions, be they relative to candidacy or validation, require sizeable sets of data.

CONCLUSIONS

Retrospective bias undermines many findings, making what is just an artefact of optimal parameter selection, interpreted as if parameters were selected at random, appear to be real. Retrospective optimization has many implicit aspects that make it generally impossible to trace and correct the bias *ab initio*. A wise countermeasure would then appear to be simply to acknowledge the existence of this bias and discard all conclusions derived retrospectively except the exceedingly unlikely ones (beyond 5 or 6 standard deviations). While sound when applied to results derived from large amounts of data, with small sample sizes, this strategy also has the unwanted effect of rejecting genuine results. With the small sample sizes often encountered in geophysics, the best strategy seems to be a two-stage approach using standard significance levels in both a first step aimed at identifying candidate anomalies, where all reasonable optimizations are allowed, and a second (forward) validation step in which no optimization is allowed.

ACKNOWLEDGMENTS

I am indebted to Ian Main and Bob Geller for many stimulating discussions and for their comments on an early draft of the manuscript. Many thanks also to Phil Stark for his efforts to refresh my statistics, and to Russ Evans for polishing the wording. This work was performed with a contribution from MURST grants 60 per cent and 40 per cent and from GNV/INGV.

REFERENCES

Anderson, P.W., 1990. On the nature of the physical laws, and 'intuition', *Phys. Today,* **43,** 9–11.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Stat. Soc.,* **57,** 289–300.

Ben-Menahem, A., 1995. A concise history of mainstream seismology: origins, legacy, and perspectives, *Bull. seism. Soc. Am.,* **85,** 1202–1205.

Bezdek, J.C., 1987. *Pattern Recognition with Fuzzy Objective Function Algorithms,* 2nd edn, Plenum Press, New York.

Cochran, W.G. & Cox, G.M., 1957. *Experimental Designs,* 2nd edn, J. Wiley & Sons, New York.

Duda, R.O. & Hart, P.E., 1973. *Pattern Classification and Scene Analysis,* J. Wiley & Sons, New York.

Eckhardt, D.E., 1984. Correlation between global features of terrestrial fields, *Math. Geol.,* **16,** 155–171.

Fisher, R.A., 1935. *The Design of Experiments,* Oliver & Boyd, Edinburgh.

Fukunaga, K., 1990. *Statistical Pattern Recognition,* 2nd edn, Academic Press, New York.

Gelfand, I.M., Guberman, S.H., Keilis-Borok, V.I., Knopoff, L., Press, F., Ranzman, E., Rotwain, I. & Sadovsky, A.M., 1976. Pattern recognition applied to earthquake epicenters in California, *Phys. Earth planet. Inter.,* **11,** 227–283.

Gonzato, G., Mulargia, F. & Marzocchi, W., 1998. Practical application of fractal analysis: problems and solutions, *Geophys. J. Int.,* **132,** 278–282.

Hsu, J., 1996. *Multiple Comparisons: Theory and Methods,* Chapman & Hall, London.

Kagan, Y.Y., 1993. Statistics of characteristic earthquakes, *Bull. seism. Soc. Am.,* **83,** 7–24.

Keilis-Borok, V.I. & Rotwain, I.M., 1990. Diagnosis of time of increase probability of strong earthquakes in different regions of the world, *Phys. Earth planet. Inter.,* **61,** 57–72.

Keilis-Borok, V.I., Knopoff, L., Rotwain, I.M. & Allen, C.R., 1988. Intermediate-term prediction of occurrence times of strong earthquakes, *Nature,* **335,** 690–694.

Knopoff, L., Levshina, T., Keilis-Borok, V.I. & Mattoni, C., 1996. Increased long-range intermediate-magnitude earthquake activity prior to strong earthquakes in California, *J. geophys. Res.,* **101,** 5779–5796.

Kossobokov, V.I., Keilis-Borok, V.I. & Smith, S.V., 1990. Localization of intermediate-term earthquake prediction, *J. Geophys. Res.,* **95,** 19 763–19 772.

Lehmann, E.L., 1986. *Testing Statistical Hypotheses,* Springer–Verlag, New York.

Main, I., 1996. Statistical physics, seismogenesis and seismic hazard, *Rev. Geophys. Space Phys.,* **34,** 433–462.

Mood, A.M., Graybill, F.A. & Boes, D.C., 1974. *Introduction to the Theory of Statistics,* 3rd edn, McGraw–Hill, Singapore.

Morelli, A. & Dziewonski, A.M., 1987. Topography of the core-mantle boundary and lateral homogeneity of the liquid core, *Nature,* **325,** 678–683.

Mulargia, F., 1997. Retrospective validation of the time association of precursors, *Geophys. J. Int.,* **131,** 500–504.

Mulargia, F. & Gasperini, P., 1995. Evaluation of the applicability of the time- and slip-predictable earthquake recurrence models to Italian seismicity, *Geophys. J. Int.,* **120,** 453–473.

Mulargia, F. & Gasperini, P., 1996. Precursor candidacy and validation: the VAN case so far, *Geophys. Res. Lett.,* **23,** 1323–1326.

Mulargia, F., Gasperini, P. & Marzocchi, W., 1991. Pattern recognition applied to volcanic activity: identification of the precursory patterns to Etna recent flank eruptions and periods of rest, *J. Volc. Geotherm. Res.,* **45,** 187–196.

Mulargia, F., Marzocchi, W. & Gasperini, P., 1992. Statistical identification of physical patterns which accompany eruptive activity on Mount Etna, Sicily, *J. Volc. Geotherm. Res.,* **53,** 289–296.

Pantosti, D., Schwartz, D.P. & Valensise, G., 1993. Palaeoseismology along the 1980 surface rupture of the Irpinia fault: implications for earthquake recurrence in Southern Apennines, Italy, *J. geophys. Res.,* **98,** 6561–6577.

Scheffé, H., 1959. *The Analysis of Variance,* J. Wiley & Sons, New York.

Simes, R.J., 1986. An improved Bonferroni procedure for multiple tests of significance, *Biometrika,* **73,** 751–754.

Stark, P.B., 1992. Inference in infinite-dimensional inverse problems: discretization and duality, *J. geophys. Res.,* **97,** 14 055–14 082.

Stark, P.B. & Hengartner, N.W., 1993. Reproducing Earth 's kernel: uncertainty of the shape of the core-mantle boundary from PKP and PcP travel times, *J. geophys. Res.,* **98,** 1957–1971.

Thurber, C.H., 1996. Creep events preceding small to moderate earthquakes on the San Andreas fault, *Nature,* **380,** 425–428.

Thurber, C.H. & Sessions, R., 1998. Assessment of creep events as potential earthquake precursors: application to the creeping section of the San Andreas fault, California, *Pure appl. Geophys.,* **152,** 685–705.

Varotsos, P. & Lazaridou, M., 1991. Latest aspects of earthquake prediction in Greece based on seismic electric signals, *Tectonophysics,* **188,** 321–347.

Varotsos, P., Alexopoulos, K. & Lazaridou, M., 1993. Latest aspects of earthquake prediction in Greece based on seismic electric signals, II, *Tectonophysics,* **224,** 1–37.

Wyss, M. & Martirosyan, A.H., 1998. Seismic quiescence before the *M7*, 1988 Spitak earthquake, Armenia, *Geophys. J. Int.,* **134,** 329–340.