Eos, Vol. 89, No. 41, 7 October 2008

A Simple Test for Changes in Statistical Distributions

PAGES 389–390

It is often difficult to determine whether a variable is undergoing a systematic change. A conventional approach to determining whether a variable exhibits a long-term change in terms of its statistical properties is to apply a trend test or some kind of time series analysis involving various degrees of complexity. Here it is shown that it is possible to indicate, through a simple test based entirely on mathematical logic, whether a series of numbers consists of independent and identically distributed (i.i.d.) numbers.

Examples of the application of the i.i.d. test include testing monthly temperatures for new record values [*Benestad*, 2004], or studying future trends in extreme monthly precipitation based on large, multimodel ensembles of climate model simulations [*Benestad*, 2006].

Conventional methods for studying changes in a variable over time often use Student's *t* tests, Kolmogorov-Smirnov or Mann-Kendall tests, or regression against time. Some of these methods assume certain types of statistical properties such as Gaussian distribution, or they involve approximations or curve fitting. Any random variable $X = [X_1, X_2, X_3,..., X_n]$ with *n* elements can be expressed in terms of its probability density function (pdf) [*Wilks*, 1995; *von Storch and Zwiers*, 1999]. Statistical analysis usually assumes that each data point is independent and that the intrinsic pdf is constant (the same for $X_1, X_2, X_3,..., X_n$), i.e., that *X* is an i.i.d. number.

Often it is essential to ensure that each data point is not affected by a subsequent value or that the pdf is not changing. An i.i.d. test [*Benestad*, 2003; *Vogel et al.*, 2001] provides an indication of whether *X* is (1) independent and (2) identically distributed (whether the pdf for *X* is changing).

For i.i.d. variables, the order of elements has no structure (i.e., the values are ordered after magnitude), and the probability that the last value in *X* is greatest can be estimated by considering the number of permutations with arbitrary element X_i being last, divided by the total number of possible permutations: $(n - 1)!/n! = 1/n$. Thus, $Pr(X_n > [X_1,...,X_{n-1}]) = 1/n$.

The i.i.d. test can then be based on the order of the elements in *X* and the occurrence of record-breaking events (i.e., new record values) (Figure 1a). The test applies equally to chronological order ("forward test") and reverse chronological order ("backward test"). It is possible to use the expression $1/n$ to describe the probability of observing new record values for an i.i.d. variable [*Benestad*, 2003; *Vogel et al.*, 2001] (Figure 1b), providing a rule for the expected number of record events in a series of *n* observations: $E(n) = \sum_{i=1}^{n}(1/i)$. If *n* is large, then $E \sim \log(n)$, and $\exp^{E(n)}$ scales approximately linearly with *n*. The actual observed number of record events, *N*, can then be compared with $E(n)$ by plotting $\exp^N$ against *n* (Figure 1c).

Examples of the application of the i.i.d. test include testing monthly temperatures for new record values [*Benestad*, 2004] or studying future trends in extreme monthly precipitation based on large multimodel ensembles such as the third phase of the Coupled Model Intercomparison Project (CMIP3) [*Benestad*, 2006].

The i.i.d. test makes no assumption about the type of pdf, as long as *X* contains no ties for the highest number (i.e., there is one maximum value, not two or more identical ones) and there is no physical upper boundary limiting the highest values of *X* [*Benestad*, 2004]. However, one argument against the i.i.d. test is that it ignores a large fraction of the information stored in the series by looking only at the record values. However, if the i.i.d. test is applied to several parallel data series, it is possible to take advantage of a greater part of the stored information content. For instance, *X* can be split into two parallel series—one representing every odd data point $[X_1, X_3, X_5,...]$ and the other representing every even data point $[X_2, X_4, X_6,...]$—and the i.i.d. test can be applied to each series. The splitting can be taken further, and the use of many such parallel series also provides a remedy for large sampling fluctuations associated with the occurrence of record events for a single series.

Confidence limits can be estimated from binomial distributions for sets of parallel series or from simple Monte Carlo simulations (Figures 1b and 1c).

A change in climate often implies a change in the pdf. When the i.i.d. test is applied to precipitation, however, it is important to distinguish between dry days and rainy days, as the rain amount has different distributions for these two categories. Furthermore, geophysical data are often not i.i.d. due to the presence of a seasonal cycle; however, subsampling by picking a selection of days at the same time of year over many years may in principle yield many parallel series, with series each holding data that are i.i.d.

The i.i.d. test is useful in extreme value modeling and return-value analysis, as a changing pdf would invalidate the extrapolation of probability based on the past. Such a test can also reveal whether analog models are flawed [*Imbert and Benestad*, 2005], as these models cannot extrapolate values outside of their historical range (i.e., the span between the lowest and highest value measured in the past). The i.i.d. test can be used to determine whether an instrument cuts off the measurement at an upper range [*Benestad*, 2004] and to test for trends in the extremes.

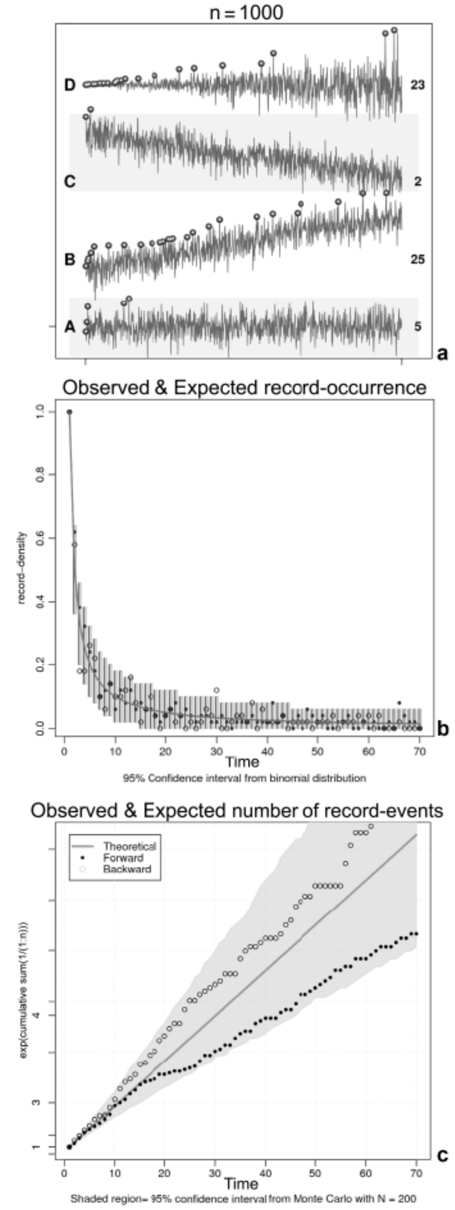

Fig. 1. (a) Hypothetical examples of record events (circles) in constructed series: graph A, i.i.d.; graph B, positive trend; graph C, negative trend; and graph D, growing amplitude. Numbers on the right of the graph indicate the number of times a new record was made. (b) Counted number of records from the M = *200 Monte Carlo simulations of case A (i.i.d.) divided by* M *(symbols), and expected fraction (grey curve). (c) The counted number of record-breaking events,* N *(symbols), as a function of* n, *with the gray line showing* E(n) *and the shaded area showing the 90% confidence interval from a set of Monte Carlo simulations. The graphs were generated with the R software (see http://cran.r-project.org) and the package "i.i.d.-test" (see http://cran.r-project.org/web/packages/iid.test/index.html).*

A rejected null hypothesis (i.e., not i.i.d.), however, does not necessarily imply that there is a substantial trend in the time series; a rejected hypothesis also could mean there is low-frequency variability. Therefore, subsamplings may be necessary for further i.i.d. tests to confirm the presence of a trend.

Although the i.i.d. test is based on simple mathematics, there is little reference to i.i.d. tests in the scientific literature. This lack of awareness may suggest neglected opportunities in the way that extremes are analyzed.

Acknowledgment

This research has been supported by the Norwegian Meteorological Institute and the NorClim project (grant 178246).

References

Benestad, R. E. (2003), How often can we expect a record event?, *Clim. Res., 25*, 3–13.

Benestad, R. E. (2004), Record-values, nonstationarity tests and extreme value distributions, *Global Planet. Change, 44*(1-4), 11–26.

Benestad, R. E (2006), Can we expect more extreme precipitation on the monthly time scale?, *J. Clim., 19*(4), 630–637.

Imbert, A., and R. E. Benestad (2005), An improvement of analog model strategy for more reliable local climate change scenarios, *Theoret. Appl. Climatol., 82*(3-4), 245–255, doi:10.1007/s00704-005-0133-4.

Vogel, R. M., A. Zafirakou-Koulouris, and N. C. Matalas (2001), Frequency of record-breaking floods in the United States, *Water Resour. Res., 37*(6), 1723–1731.

von Storch, H., and F.W. Zwiers (1999), *Statistical Analysis in Climate Research*, Cambridge Univ. Press, New York.

Wilks, W. S. (1995), *Statistical Methods in the Atmospheric Sciences*, 467 pp., Academic, Orlando, Fla.

—Rasmus E. Benestad, Norwegian Meteorological Institute, Oslo; E-mail: rasmus.benestad@met.no

NEWS

Seismogram “Picking Error” Experiment

PAGES 390–391

The Seismogram Picking Error From Analyst Review (SPEAR) project is looking for input from analysts with experience in picking earthquake or other seismic data. The project, begun in February 2008 as a grassroots effort among the authors’ institutions, will aid the seismological community in producing more reliable results for hypocenter locations and in establishing Earth models that are more accurate. The project could develop into a continuous forum for the seismological community to better understand the errors, biases, and mental processes or steps involved in picking seismic data, and to establish standards for how to pick and assess seismic data.

Picking earthquakes entails reviewing seismograms generated from continuous monitoring for movement of the Earth’s surface. Seismic sources produce seismic phases (waves associated with characteristic arrivals, motions, and velocities) that can be distinguished from ambient noise (tidal interactions, atmospheric changes, lightning strikes, or seasonal variations in seismograms) or cultural noise (vehicular vibrations, construction, and so forth). Once a signal has been identified as a seismic source, the seismic phases can be “picked” as to where they begin along the timescale of the seismogram.

The most recent study of errors from analyst-reviewed seismograms (H. W. Freedman, *Bull. Seismol. Soc. Am., 56*(2), 593–604, 1966) found that individual analysts maintain consistency between tests—administered at different occasions—of identifying and picking the onset of an earthquake from the same seismic record. For different analysts studying the same record, Freedman found a standard deviation of 0.2 seconds for earthquakes with impulsive onsets (a sharp or pointed transition from the noise of a seismogram to the signal of a seismic source). One expects a larger standard deviation for emergent picks (which have no clear transition from noise to seismic source signal other than frequency or amplitude).

We propose to revisit this human aspect of 21st-century seismology, since most recent picking error research has focused on the reliability of automated systems. We also want to revisit picking error because Freedman’s study was conducted on the standard at that time, paper records. The seismological community has since transitioned into using digital records and associated computer programs (Dbpick, SAC, Earthworm, and Seisan are among the publicly available programs) to pick and analyze seismograms. Yet Freedman’s results still are relevant: The results pointed out that the precision of the instrumentation in that era already exceeded the training of the analysts. Since then, no additional training or standards have evolved to aid the seismological community in determining appropriate measurement practices. In Freedman’s day, the measurement error was trivial relative to the much larger error introduced by the model. Today, however, model errors and measurement errors are of similar magnitude.

The study of picking errors is hindered because most events are reviewed and studied only by local or regional seismic centers proximal to the epicenter. Thus, only a limited number of individual “pickers” can be compared. Also, most earthquake catalogs only publish the time, date, latitude, longitude, depth, and magnitude of an earthquake; the picks are archived at the source of the hypocenter solution. Without access to repeated measurements of seismic phases or picks, an error model cannot be established.

The catalog of the International Seismological Center (ISC), in Thatcham, United Kingdom, is an exception to the traditional earthquake catalog; hypocenters as well as picks, often from investigators at more than one institution, are compiled. From the pick information available in the preexisting ISC databases that were provided to SPEAR, three results were derived: A single analyst can pick repeated earthquakes very precisely; a “picking culture” or bias between different institutions is evident; and a controlled picking experiment is needed to clearly identify an accurate picking error model (C. P. Zeiler and A. Velasco, manuscript in preparation, 2008). This brings us back to Freedman’s method of, and the need for, multiple seismologists—representing different levels of experience—to pick a common data set.

Thus, we need a variety of analysts representing different institutions, years of picking, and experience with different seismic networks. Those interested in assisting with the SPEAR project can find instructions on the project home page (http://www.geo.utep.edu/pub/SPEAR/SPEAR.html). Volunteers pick two sets of seismograms that have different characteristics: The first picking attempt is an initial pass with no filtering or enhancement to the seismograms; in the second attempt, volunteers use their standard picking method. With the assumption that all seismic stations are equally reliable, volunteers will pick the seismograms as if they were going to solve for a location. Volunteers are requested to disclose their picking history and any information or tactics that aid them in the picking process.

The process should take no more than 1 hour for volunteers to complete. Volunteers have the option of receiving a personal picking profile (privacy will be maintained) that illustrates a volunteer’s comparative statistics.

We anticipate SPEAR to be the start of a phase-picking forum to discuss issues related to picking seismic signals. As members of the community join this collaborative effort, we will add more seismograms to address other issues that arise.

—Cleat P. Zeiler, Aaron Velasco, and Nicholas E. Pingitore Jr., Department of Geological Sciences, University of Texas at El Paso; E-mail: spear@geo.utep.edu; and Dale Anderson, Pacific Northwest Laboratory, Richland, Wash.