Journal of Scientific Exploration, Vol. 16, No. 4, pp. 571–577, 2002 0892-3310/02

COMMENTARY

Was There Evidence of Global Consciousness on September 11, 2001?

JEFFREY D. SCARGLE

Space Science Division, NASA Ames Research Center, MS 245-3, Moffett Field, CA 94035-1000
jeffrey@cosmic.arc.nasa.gov

Abstract—This note critically reviews the methodology of the accompanying papers by Roger Nelson and Dean Radin, emphasizing a key limiting feature of the experimental procedure. I personally disagree with the former's conclusion that anomalous effects have been unequivocally established. The latter's paper, analyzing the same data, views its results as suggestions to be tested using future data, which for reasons discussed below is the only possible result of exploratory analysis. While I judge the degree of cogency of all of the results in both papers as low, this note is essentially a set of suggestions that I hope will encourage both you, the Reader, to judge for yourself and the researchers in this field to improve their methodology.

Keywords: statistical significance—cumulative distributions—exploratory analysis

Introduction

The two papers "Coherent Consciousness and Reduced Randomness: Correlations on September 11, 2001," by Roger D. Nelson (hereafter RDN) and "Exploring Relationships Between Random Physical Events and Mass Human Attention: Asking for Whom the Bell Tolls," by Dean Radin (hereafter DR), describe attempts to detect possible effects of reactions by many persons to the singular events of September 11, 2001, on the random number generator system comprising the Global Consciousness Project (GCP). In summary, both papers search for possible departures from the randomness to be expected were there no mind-matter influences. Radin's is essentially an exploratory analysis, while Nelson's goal is to test predefined hypotheses.

Statistical Issues

You, the Reader, will and should judge the claims in these works for yourself. I wish to point out a few data analysis issues that might aid in this assessment.

This is not meant to be a complete analysis, and in particular I do not cleanly separate the two papers, which are closely related and interdependent in some ways. I do attempt to separate my opinions on issues that are controversial from those generally agreed on by the scientific data analysis community.

Throwing Out the Baby With the Bath Water?

GCP data processing includes the application of a logical XOR operator to the bit stream. The process actually involves an XOR between two physical random digit streams, followed by a deterministic flipping of every second bit (Roger Nelson, personal communication). The purpose of this operation is to filter out "... trends attributable to spurious physical sources" (RDN) and "to ensure that the mean output is unbiased regardless of environmental conditions, component interaction, or aging" (DR).

However, the bit flipping operation also renders the GCP completely insensitive to a whole class of possible effects. For example, suppose there were a mental signal—perhaps transcending ordinary human senses and known laws of physics—generated by, and acting coherently in, groups of humans. Suppose further that this signal acts to change the relative frequency of 0s and 1s in RNG's by a statistically significant amount. Isn't this what global consciousness is all about? No, according to GCP! The GCP system is insensitive to such a signal because the bit flipping operation would null it out (along with possible interference). The GCP is seeking evidence of effects that operate directly on the "final answer." Different Readers will no doubt have different assessments on this matter.

Perhaps expressing my personal astonishment at all this, I characterize what GCP is seeking as hyper-transcendental—i.e., the system is purposefully sensitive to only effects that transcend both direct sensory detection and elementary causality as described above. The GCP explicitly excludes direct coherent effects which, if discovered, would revolutionize science in a heartbeat. Their position seems to be that such "physicalist" causal effects are not being pursued because they have already been ruled out (RDN, personal communication).

Z-Scores and p-Values

Many authors point out the dangers in the use of Z-scores, often called p-values, for this kind of analysis (e.g., in the precise context relevant here, and in this Journal; see Jefferys 1990; Sturrock 1994, 1997). Is this demanding a higher standard than is normally acceptable? There may be something to the idea that exceptional claims require exceptional evidence, but that is not my point—rather that everyone seeking Truth should use correct statistical analysis, and the entire community needs education.

Overestimation of statistical significance of what are really random fluctuations due to the use of p-values may well be operative in these analyses.

However, I will not overemphasize this point, since the authors face the following dilemma, which is perhaps the reason neither refers to the above three absolutely on-point articles: These statistical issues are, unfortunately and inexcusably, still not well understood by the scientific community. Many would look on a Bayesian analysis with suspicion, no matter how proper and straightforward. Nevertheless, I believe that this kind of result will not be generally accepted unless both Bayesian and classical p-value analysis agree, and both show the same anomalous effects.

Cumulative Distributions?

While RDN's Figures 1 and 2 are impressive on first glance, there are several considerations that in the end make them less so. Note that most authors in this subject examine a specific statistic, χ^2 or something closely related. This is one way to examine possible departures from pure randomness, but it is not the only way. Both papers discuss the computations, and DR's list of steps in the section Variance Analysis is particularly explicit. Two quantities are displayed in various places, essentially cumulative sums or running means of χ^2.

Cumulative χ^2 is a very tricky quantity. Right off the bat the χ^2 statistic throws away the sign of the effect, thus limiting its scope. More important, the summation operation has the effect of producing coherent structure where there is none. To see this the Reader need only plot cumulative sums of any zero-mean random number. Technically, this operation turns white noise (flat power spectrum) into correlated noise, sometimes called $1/f$ noise because the power spectrum has this form. The coherent structure always seen in such plots is the result of the memory introduced into the time series by the cumulative summation and has nothing to do with any structure in the original time series. The phrase ''cumulative deviation'' suggests that nice idea of trying to accumulate the effects of a small putative signal, but unfortunately this statistic also accumulates statistical fluctuations in such a way that plots of it have the visual appearance of coherent structure.

Running means have a similar effect. The only statistically safe procedure is to use independent means (i.e., averages of blocks of the data that do not overlap each other), for then the resulting values are statistically independent of each other, making it relatively straightforward to judge statistical significance. Use of running means with overlap introduces the same kind of bogus structure as does cumulative summation.

Actually, both of these operations are essentially different versions of the same thing. Cumulative sums are like a running mean where the width of the block of data being averaged varies, such that there is maximal overlap between different blocks. Both methods are sensitive to choices, such as where the summation begins, the size of the running mean window and its degree of overlap. The Reader can verify that slightly different choices for these make very big differences in the appearances in many of the Figures in these papers.

For all of these reasons a better statistic for hypothesis testing is the independent running mean, with careful control on the choice of the averaging window size. The Reader can experiment with this statistic. My experiments yielded only plots that have the appearance of white noise, over a broad range of smoothing values.

Data Fiddling

Both authors are aware of the pitfalls of *post facto* analysis of random data. For example, DR states that "... the results will be useful primarily in developing future hypotheses." A careful discussion of this crucial issue is in order.

It is well-known[1] that exploratory data analysis can almost always yield spurious structure that appears significant, even after accounting for the random observational errors. I don't know of any commonly used name, so I will call this process data fiddling. The problem is not just systematic errors, but the fact that it is difficult or impossible to properly account for the extra degrees of freedom generated by fiddling with various aspects of the data. The underlying reason is that one in effect performs many experiments which are almost always interdependent and dependent on the data in ways that are difficult, if not impossible, to account for statistically.

For example, RDN finds a result in Figure 1 that is not very significant, so he looks at more data to yield the apparently more significant result in his Figure 2. I do not object to this examination of the data in "the larger context," but do believe that it should be accompanied by the comment that at this step one is going outside the scope of a pre-defined hypothesis and performing exploratory analysis.

"Formal" vs. Exploratory Analyses

Correctly, researchers in this area address the problems of data fiddling by dividing their research into two phases: an exploratory one, where one examines data in a rather free way in order to frame hypotheses to be tested in the later phase, where completely defined hypotheses are tested against new, independent data.

The DR paper is entirely exploratory, but RDN attempts to be entirely formal[2] in that this paper directly tests predefined hypotheses with no data fiddling. Discussion and lists of such hypotheses are in a Prediction Registry at http://noosphere.princeton.edu/ (anyone can retrieve much of the RNG data from this excellent site). This is a good idea, but I feel its implementation falls short in several respects.

A prediction must be specific and complete enough that testing it is a completely objective matter, involving no choices. Otherwise one is back in the arena of exploratory analysis, at least partially. In a nutshell, testing should be achievable by turning over the data to a pre-written computer program which applies a fixed method of analysis with no free parameters. I do not believe that any of the predictions in the registry, or "formal analyses" carried out on the

GCP data, strictly satisfy this criterion. The Reader may wish to consult the prediction registry for the September 11, 2001, events and judge whether an unambiguous hypothesis is there framed. In my opinion, the comments there merely define a general time frame, and leave much fiddle room.

Furthermore, the predictions often seem to reflect the predictor's intuition about how the anomalous effects should behave (e.g., Dean Radin is quoted thusly, with regard to the events under discussion: ''I'd predict something like ripples of high and low variance, as the emotional shocks continue to reverberate for days and weeks.'' Such predictions should be accompanied by a statement of origin—was it suggested by previous exploratory analysis, a gut feeling, intuitive guess, or whatever? The registry does add, in reference to the above quote: ''Although this is not sufficiently specific to be included in the formal database, it is effectively a prediction that the variance around the time of the disaster should deviate from expectation. At least I wish to make that more specific version of the prediction, with medium confidence, seconds resolution.'' One presumes that this is then the ''formal prediction.'' (I do not understand why the confidence of the prediction is relevant.) However, the variance might behave in many ways, and without specifying a test and an algorithm for implementing it, there is much fiddle room indeed.

Skating on the Edge of Statistical Significance

Very often in science when a small effect is suspected, a few more experiments with improved technology and more data quickly settle the question, either validating or discrediting the suspicion. In cases where there is no effect present, there can be a long period where analysis of new data is constantly modified to tease out small effects using data fiddling, and in the end there is no effect present. One gets the impression that RNG studies of the kind discussed here and previously have been skating on the edge of statistical significance for many years. Why have none of the putative effects found earlier become of undisputable signal-to-noise with the advent of tremendously greater volumes of data? Here I do not believe to be valid the argument that RDN used to deflect a similar, but different, criticism—namely, why didn't the September 11 events show up indisputably, since it was so much more mentally anguishing than other events that were claimed to show small effects? RDN's argument is something like the following: the effects need not be linearly additive. I have never heard anyone question the applicability of $\frac{1}{\sqrt{n}}$ statistical improvement, which indeed seems to be a primary motive behind the expansion of the GCP RNG network.

Conclusions

Because it is meant to be exploratory, DR in my opinion cannot be faulted for making unsupported claims. This exposition has a pleasing, almost tintinnabular style that should resonate with many readers.

Setting aside my opinion that the separation of *a priori* from *a posteriori* discussion could be improved, does the analysis for the former kind in RDN justify the following claims?

"... unmistakable structure in data that should be genuinely random,"
"... definitely show anomalous deviations ...," and
"... the results of our analyses are unequivocal."

In my opinion these claims are unwarranted, because the putative *a priori* hypotheses are not sufficiently precisely framed prior to the analysis, and because the reported statistical results, even taken at face value, are unconvincing.

Here are some recommendations for future work that derive from the above discussion.

- Carry out exploratory analysis in a more open-minded way. In the current context, seek any departures from randomness, not just a limited class of effects that are deemed likely and/or are easy to deal with.
- Do not perform the XOR operation, work to improve shielding of spurious interference, or—if systematic errors are really suspected—record the data both with and without XOR.
- Implement more discipline on "formal" predictions: to eliminate data fiddling, ensure that they are complete and specific enough to be implementable automatically, with no human intervention via selection or adjustment of parameters.
- Carry out all tests in the corresponding automated way. Most convincing would be blind, parallel analysis of suites of time series: data with no signal is mixed with the data under investigation.
- Use straightforward Bayesian analysis in lieu of, or in addition to, discredited classical "p-value" tests.

I believe that this subject is still completely in the exploratory phase, and that none of the results discussed in either of the two subject papers is compelling in any degree.

Acknowledgements

I am grateful to both authors for helpful discussions, and especially to Roger Nelson for assistance with the data. I applaud the current authors, and others in this field, for those steps they have taken toward applying a rigorous statistical methodology, and especially for having a very open web site with raw data and many methodological details. Of course, a critical analysis such as that presented here would be much more difficult if the GCP were not as open as it is. But I encourage the further steps recommended here. Finally, I applaud RDN's inclusion of the May and Spottiswoode reference; this paper reaches a negative conclusion on the same data.

Notes

[1] Well known, but not always remembered. In perhaps the largest statistical blunder in the astronomical literature (Varshni 1976), *a priori* statistical analysis was applied to an *a posteriori* situation, resulting in an overestimate of the significance of clustering in quasar redshifts by many tens of orders of magnitude, and reaching the conclusion that the Earth is the center of the Universe! See Weymann et al., 1978, for details.

[2] My feeling expressed in the previous section—that there are elements of the discussion that are exploratory in nature and should be labeled as such—is probably minor compared to the issues discussed here.

References

Jefferys, W. H. (1990). Bayesian analysis of random event generator data. *Journal of Scientific Exploration, 4,* 153–169.

Sturrock, P. A. (1994). Applied scientific inference. *Journal of Scientific Exploration, 8,* 491–508.

Sturrock, P. A. (1997). A Bayesian maximum—entropy approach to hypothesis testing, for application to RNG and similar experiments. *Journal of Scientific Exploration, 11*, 181–192.

Varshni, Y. P. (1976). The red-shift hypotheses for quasars: is the Earth the center of the universe? *Astrophysics and Space Science, 43,* 3.

Weymann, R., Boroson, T., & Scargle, J. (1978). Are quasar redshifts randomly distributed?: comments on the paper by Varshni "Is the Earth the Center of the Universe." *Astrophysics and Space Science, 53,* 265.