LA-14453

Approved for public release

Distribution is unlimited

Determining the significance of associations between two series of discrete events: bootstrap methods

Los Alamos National Laboratory, an affirmative action/
equal opportunity employer, is operated by Los Alamos
National Security, LLC, for the National Nuclear Security
Administration of the U.S. Department of Energy under
contract DE-AC52-06NA25396.

LA-14453-MS
Issued: January 2012

Determining the significance of associations between two series of discrete events: bootstrap methods

J. T. Niehof
S. K. Morely

Los Alamos
NATIONAL LABORATORY
EST. 1943

Determining the significance of associations between two series of discrete events: bootstrap methods

J. T. Niehof S. K. Morley

January 26, 2012

Abstract

We review and develop techniques to determine associations between series of discrete events. The bootstrap, a nonparametric statistical method, allows the determination of the significance of associations with minimal assumptions about the underlying processes. We find the key requirement for this method: one of the series must be widely spaced in time to guarantee the theoretical applicability of the bootstrap. If this condition is met, the calculated signifance passes a reasonableness test. We conclude with some potential future extensions and caveats on the applicability of these methods. The techniques presented have been implemented in a Python-based software toolkit.

1 Introduction

A question of frequent interest across many fields is the possible relationship between two types of observation, potentially with some time delay. Several familiar tools, such as regression analysis and cross-correlations, relate continuously varying quantities. We expect, for example, a strong correlation between the time series of temperatures in White Rock and Los Alamos. A somewhat weaker correlation would be evident between Los Alamos and Oklahoma City, with Los Alamos expected to lag by about half an hour in the dominant diurnal variation.

Less familiar are tools for associating point events, rather than continuously varying quantities. For instance we may be interested in a possible association between hailstorms in Los Alamos and work requests at body shops in northern New Mexico. Although potentially reducible to pseudo-continuous quantities (e.g., storms per month and body shop jobs per month), this reduction loses the temporal association between individual events and may require a period of averaging longer than the delay between types of events. This report reviews and extends techniques for determining the level of association between two types of event, any delay in the association, and a confidence that the observed level of association is significant (beyond that expected from chance.) These techniques are agnostic with respect to causality and work even if only a subset of each event class is associated with the other: there may be reasons other than hailstorms to seek body shop work. Subject to the caveats in section 3, they also work for clustered events, e.g., multiple requests for body work following a single hailstorm.

This report builds on methods presented in the space sciences by *Morley and Freeman* [2007], who examined the association between magnetospheric substorms and "trigger events" in the upstream solar wind. Python code implementing the methods of this report is available in the LANL SpacePy package [LA-CC-10-064; *Morley et al.*, 2011], available under an open source license from `http://spacepy.lanl.gov/`. These techniques are applicable to a range of scientific studies.

2 Review of techniques

2.1 Association analysis

We are concerned with two time series of events, called series A and series B throughout this manuscript. Each series is described by a list of times when events of the appropriate type occurred. Each event of type A is considered identical to all other type A events, and type B events are similarly identical. We wish to determine whether events of type B happen more frequently in some period of time before or after events of type A; that is, whether there is some association between series A and series B. *Cox* [1955] developed these methods, here called "association analysis," relating textile defects to halts of weaving machines. They were further developed by *Brillinger* [1976], in the context of neuronal firing. Figures 1 and 2 of *Brillinger* [1976] provide an excellent example of the conversion from a continuously varying measurement to a series of events. *Morley and Freeman* [2007] provide another example: converting from a time series of solar wind magnetic field measurements to a list of "trigger" event times for one process, and (implicitly) from a time series of various geomagnetic indicators to a list of substorm times. The presentation here is after *Morley and Freeman* [2007].

We consider a length of time T in which we have measurements of a series of N_A type A events, where the i^{th} event is denoted a_i, and N_B type B events, where the j^{th} event is denoted b_j. We assume a time lag u between the series of events, and a half-window h representing an uncertainty in the timing of events or in the delay between them. Then the i^{th} individual association on series A is defined:

$$c_{A,i} = \#\{|b_j + u - a_i| < h\} \tag{1}$$

where $\#\{\}$ is the number of events in set $\{\}$. In the zero-lag case, this is the number of B events b_j which fall within half-window h of a particular A event a_i. A non-zero lag has the effect of time-shifting series B with respect to series A. Figure 1 presents a schematic example. Note that, in this formulation, overlapping half-windows do not add to the individual associations: one event must fall within the half-window of the other.

The total association number, usually just called the association number, as a function of lag and half-window is:

$$n(u,h) = \sum_{i=1}^{N} c_{A,i} \tag{2}$$

Note that the association number is independent of the choice of which series is A and which is B (reversing the series simply reverses the order of summation), but the individual associations c are not. Section 3 discusses the consequences of this asymmetry. The association number can be viewed as "the number of occurrences where an A-type event and a B-type event are within a half-window of each other", where, e.g., two A events and two B events all falling within a single half-window would result in four such associations.

In practice we often have no a priori knowledge of the lag between events, but wish to determine it. Thus the association number is calculated by equation 1 for a large number of possible lags u covering the expected range of lags between the processes, a computationally intensive process which nonetheless is tractable and even fast on modern computing equipment. The half-window h can also be optimized through a search procedure (section 4).

The result is a list of $n(u,h)$ generally plotted for a fixed h, e.g., figure 2 (produced by the PoPPy module of SpacePy). This result is from a pair of synthetic series. Series A is a series of 360 events whose times are determined by drawing from a uniform random distribution on the interval

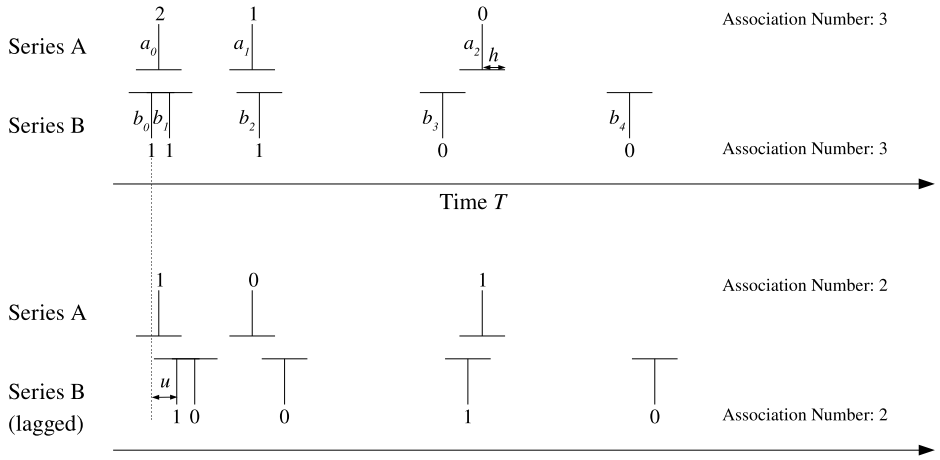

Figure 1: Two series of different types of events (A and B), represented as vertical segments on a timeline (a_i, b_j). Each event is labelled with its individual association $c_{A,i}$ or $c_{B,j}$, the number of events of the other type within a half-window h of the event. Top is the zero-lag case; bottom, with series B shifted by lag u.

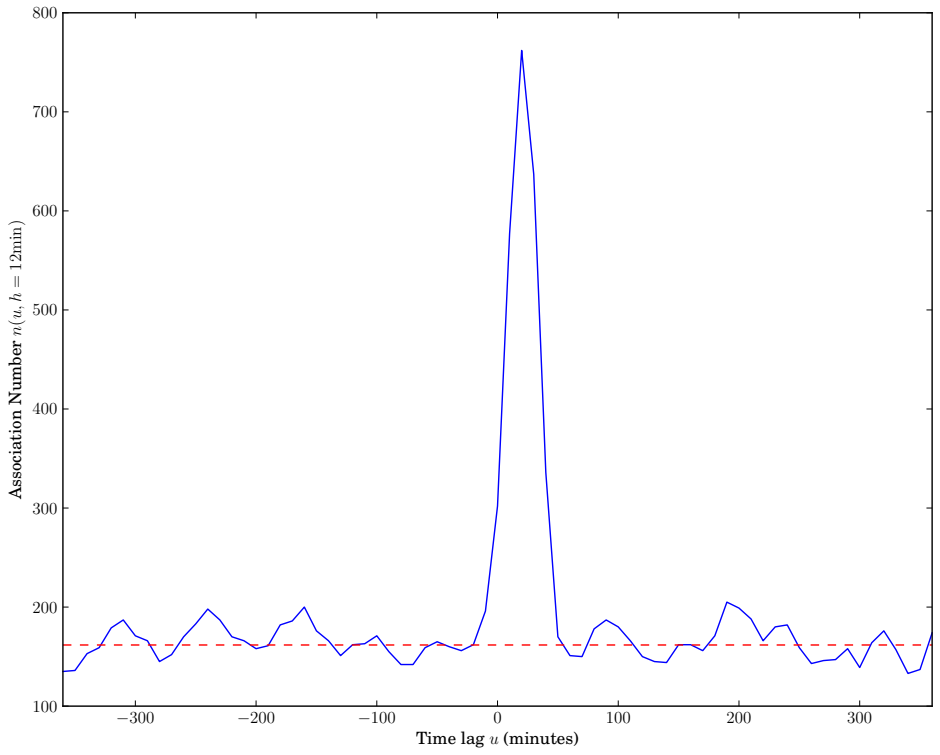

Figure 2: Association number as a function of lag for a synthetic example.

[-90, 630] and using the value as the event time (here given units of hours). Series B is a set of 882 events for which the times were determined using a Gaussian probability of occurrence around each event in series A, limited to the interval [0, 540]. The peak of the Gaussian is 20 minutes after the series A events, the half-width 10 min. Thus, as expected, the association number increases around lags of 20 min., i.e., with series A preceding series B by 20 min. The window half-width for this analysis is 12 min.

Several methods can be used to distinguish "real" associations from chance. The dotted red line in figure 2 shows one method. At very high lags, we expect any physical association between the series to be broken. Thus the association number at these high lags should tend to that of chance association; this is called the aysmptotic association number $n(u \to \infty, h)$. "Very high" depends on the nature of the problem; in this case, with a known 20 min. offset and half-width of 10 min., 100 min. should suffice. The line plotted is the median of the association numbers from the 0^{th} to 20^{th} percentile of the lags and from the 80^{th} to 100^{th} percentile. This asymptotic association number is about 162. Association numbers above this line indicate an association greater than chance; below, an association less than chance, implying a reduction in occurrence of one set of events at some lag relative to the other set. The association analysis says nothing about causality: although the series plotted shows an association between B events and A events 20 min. later, B does not necessarily cause A; some external process may drive both. It would be unexpected for A to somehow cause B in this instance.

An alternative means of finding the "chance" association number is to examine the expected level of association if events were randomly distributed [cf. *Brillinger*, 1976]. In this case, the probability per unit time of a B event is:

$$p_B = \frac{N_B}{T} \tag{3}$$

Then the number of type B events within a half-window h is $2hp_B = \frac{2hN_B}{T}$, and this number of events is expected to be associated with each A event. The predicted total association number, summing over all A events, is then:

$$n_{expected}(u, h) = \frac{2hN_AN_B}{T} \tag{4}$$

For the pair of series presented, this evaluates to 166, close to the asymptotic number of 162.

There are two subtleties to equation 4: first, to justify equation 3, B events must have a constant probability over time and be independent of each other, i.e., B must be essentially a Poisson process. As derived, there is no such constraint on series A; thus, by the symmetry of the derivation, either series A or B must be Poisson. Otherwise, the exact value of $n_{expected}$ may be inaccurate, but we would still expect scaling proportional to N_A, N_B, and $\frac{h}{T}$. The second subtlety is an implicit u dependence: as u varies, the shifted series B may not completely overlap with series A, and so in practice N_A, N_B, and T apply to the region of overlap, varying slightly with u. In the extreme case where all events in series A occur well before any event in series B, there are no events in the overlap and $n_{expected}$ is zero.

In addition to the expected strong peak at 20 min., figure 2 shows several minor peaks, e.g., at $u = 190$ min., implying a positive association. This lag is well into the asymptotic region where no association is expected: clearly there is "noise" on the association number and not every peak is significant. One approach to identifying significant peaks is to take the asymptotic variability as the uncertainty and only identify peaks of greater magnitude than those in the asymptotic region;

this was suggested by *Brillinger* [1976], who also computed more robust confidence intervals based on the assumption of Poisson processes. Section 3 computes confidence intervals for an association analysis with much weaker requirements on the input series, relying on a nonparametric statistical tool: the bootstrap.

2.2 Bootstrapping

Any set of measurements on a system represents a sample of a population, and a metric derived from that sample represents an estimate for that metric over the population. Knowledge of the range of likely values for the metric on the population provides more information than this single value. Usually two assumptions are made to infer the population statistics from the sample: first, the sample is assumed to be representative of the population; second, the distribution of the population is assumed (frequently Gaussian or Poisson.) If there is no a priori reason to assume a well-behaved population, most statistical methods are not justified. These limitations are not always heeded in publications.

Efron [1979] noted that, if the sample is truly representative of the population, this is sufficient to reconstruct the population statistics in a computationally intensive way. *Diaconis and Efron* [1983] provide a simple explanation of this "bootstrapping" process: given a sample of size N, randomly select N values from this sample with replacement to create a "surrogate" sample. By resampling with replacement, some values from the original sample may appear multiple times in the surrogate; others may not appear at all. Repeating this resampling many ($\gtrsim 10^3$) of times, and each time calculating the desired metric from the surrogate, will provide a sample of possible metrics that reflect the statistics of the underlying population. Then, for example, the 95% confidence interval for the metric on the population can be inferred from the 2.5^{th} percentile and 97.5^{th} percentile of the surrogate metrics.

As an example, consider the series 1, 2, 5, 5, 6. The sample mean is 3.8; what is the range of likely means for the population? The standard deviation is 1.9, so assuming a normal population, the 95% confidence interval (2σ) for the population mean is 2.1 – 5.5. But the sample is clearly not normal, and we have no reason to expect the population would be. Applying the bootstrap, two possible surrogates are 1, 1, 5, 5, 5 (mean 3.4) and 1, 5, 5, 1, 6 (mean 3.6). For a series of length N, with N members of the series to choose and N possibilities for each, there are N^N possible surrogates, 3125 for this small sample. Generation of 1000 bootstrapped surrogates yields an estimated 95% confidence interval of 2.0–5.4.

The requirement of a representative sample is not sufficient for a trustworthy bootstrap. *Singh* [1981] strengthened the theoretical justification for the bootstrap but demonstrated its failure for dependent data: that is, where a value in a sequence has some dependency on the preceding value or values. *Künsch* [1989] and *Liu and Singh* [1992] developed a modified version, the "moving block bootstrap" (MBB), for dependent data.

The MBB works by resampling blocks of values in the sample, rather than individual values. For a sample of size N, overlapping blocks of M samples each are taken, resulting in $N - M + 1$ blocks, each block shifted by one sample from the previous. These blocks are resampled into a new surrogate series, and the desired metric calculated on the surrogate. Multiple variations on the MBB have been proposed to counter the obvious "edge effects" (the first and last value of the sample only occur in one block each); *Lahiri* [2003] references several examples but demonstrates that the simple MBB provides accuracy comparable to the more complicated methods. *Lahiri* [2003] also addresses the question of optimal block size; a rule-of-thumb is that the block should

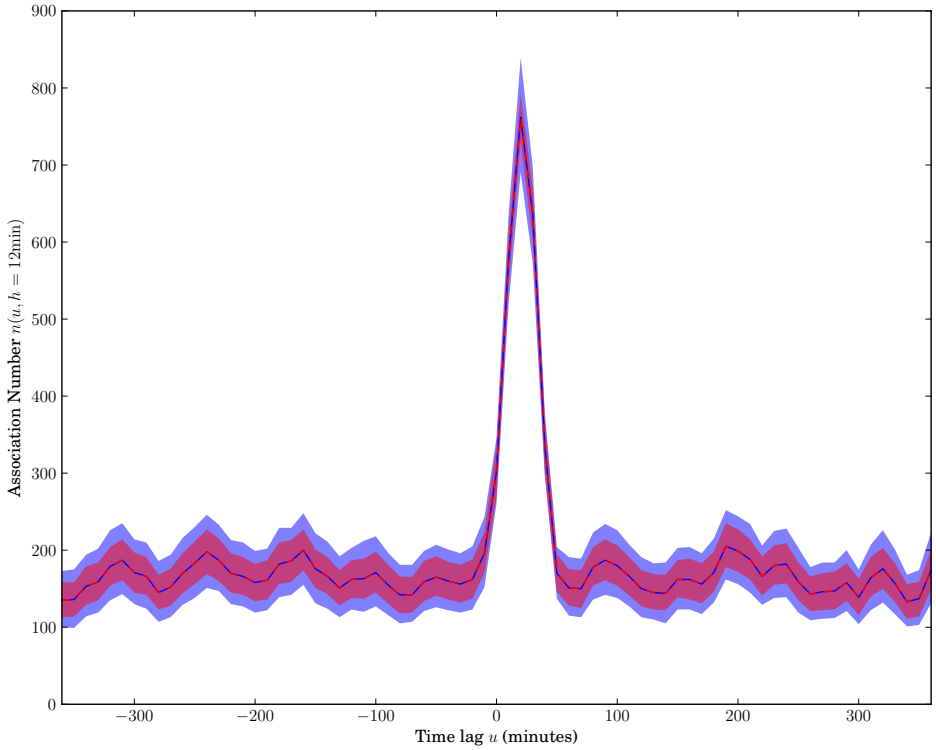

Figure 3: Association number as a function of lag, with 95% confidence intervals. Blue line and blue fill are association number and confidence interval based on calculating individual associations for series A, then summing; red line and red fill are based on summing over individual associations for series B. Areas of overlapping confidence intervals are magenta.

be of order $N^{\frac{1}{4}}$ to $N^{\frac{1}{3}}$, but far more involved algorithms are available.

All bootstrap methods are oriented towards a single series of values. The bootstrap function in SpacePy accepts such a series and the function to apply to the surrogates, returning the requested confidence interval and, if desired, the probability that the true value is below some reference. The following section adapts these methods to association analysis, where there are two series of discrete events occurring at a particular time, rather than with a particular value.

3 Confidence intervals for associations

Morley and Freeman [2007] constructed bootstrap confidence intervals on the association number by resampling the individual associations, $c_{A,i}$ (equation 1), and evaluating the appropriate percentiles of the sums across the surrogate series. This approach breaks the symmetry between series A and B, as the series $c_{A,i}$ is likely to be very different from $c_{B,j}$, although the association numbers are the same. Figure 3 shows the data of figure 2 with 95% confidence intervals calculated over 4000 bootstrap iterations. The calculation was performed twice: once as in figure 2, and once with series A and B swapped. In the former case the bootstrap resampled over the individual asociations for series A; in the latter, over those for series B. For the calculation with swapped series, the signs of the lags were inverted, as A leading B is equivalent to B lagging A. It is clear that resampling

over the 360 individual associations for series A yielded more conservative confidence intervals than resampling over the 882 series B individual associations.

The question we wish to answer is symmetric: what is the probability that the observed association is attributable to chance? The confidence intervals based on association in one direction must be more "correct" than the other, giving a better prediction for the possible range of the population. It is tempting to ascribe the tighter confidence interval when resampling over B to $\frac{1}{\sqrt{N}}$ sampling effects, and indeed increasing the density of events in series B without changing the distribution does reduce the confidence interval. This only increases the disparity between the confidence interval calculated by bootstrapping on series A (which does not change) and bootstrapping on series B. Furthermore, it is possible to create a pair of series such that bootstrapping on the shorter series will result in a narrower confidence interval. Clearly some means must be found to select the appropriate series to bootstrap on.

Since the bootstrap requires the independence of individual values in the sample, what if some dependence between events violates this assumption? To test this possibility, we implemented a version of the moving block bootstrap. The MBB defines a block as containing a certain number of values from the sample. We chose to define a sample as a single coincidence between series A and series B, with an occurrence time halfway between the time of the series A event and the (lagged) time of the series B event. Each block was then defined as a period of time containing a certain number of coincidences, with block boundaries halfway between coincidence times. Each block then had a coincidence count $N_{A,B}$, time T, and counts of A, B events N_A, N_B. Blocks were resampled with replacement to construct surrogates, and totals of $N_{A,B}$, T, N_A, N_B were computed by summing over the blocks included in the surrogates. The number of blocks in each surrogate was chosen to match the total T to that of the original series. $N_{A,B}$ for each surrogate was scaled by the ratio of the expected value (equation 4) for the original, nonresampled series to the expected value for the resampled series. The confidence intervals from the resulting bootstrap were completely symmetric when series A and B were swapped.

Figure 4 shows the MBB-computed 95% confidence intervals after 4000 bootstrap iterations compared to the method of figure 3. The MBB has much better agreement with bootstrapping over the series A individual associations than bootstrapping over the series B associations, particularly at the central peak. The MBB estimate appears to be somewhat asymmetric in the asymptotic region, with the center of the 95% confidence interval at higher association numbers than the sample association number. In the asymptotic region (absolute lags >= 120 min.), the fraction of the confidence interval above the calculated association number is larger than the fraction below by 17% (mean) to 18% (median) of the total confidence interval width. In contrast, asymmetries from bootstrapping on the individual associations are of order 2%–3%. There are at least three possible reasons for this offset. With low association numbers, there are few coincidences and thus small blocksize and/or small number of blocks to bootstrap over, and the sample may not be large enough to meet the bootstrap's requirement that it be representative. It is also possible that the requirement of blocks containing a certain number of coincidences biases the estimate. Finally, this implementation of the MBB is based on the expected association total, which lies in the interval 164.8–167.4 depending on the lag; the calculated asymptotic association number (based on the median of largest lags, positive and negative) is 161.7. Nonetheless, we can with reasonable confidence conclude that bootstrapping over the series A association numbers provides better confidence intervals than the series B association numbers. This case mirrors that of *Morley and Freeman* [2007] and justifies their bootstrapping approach of modeling the sampling variation in the individual associations.

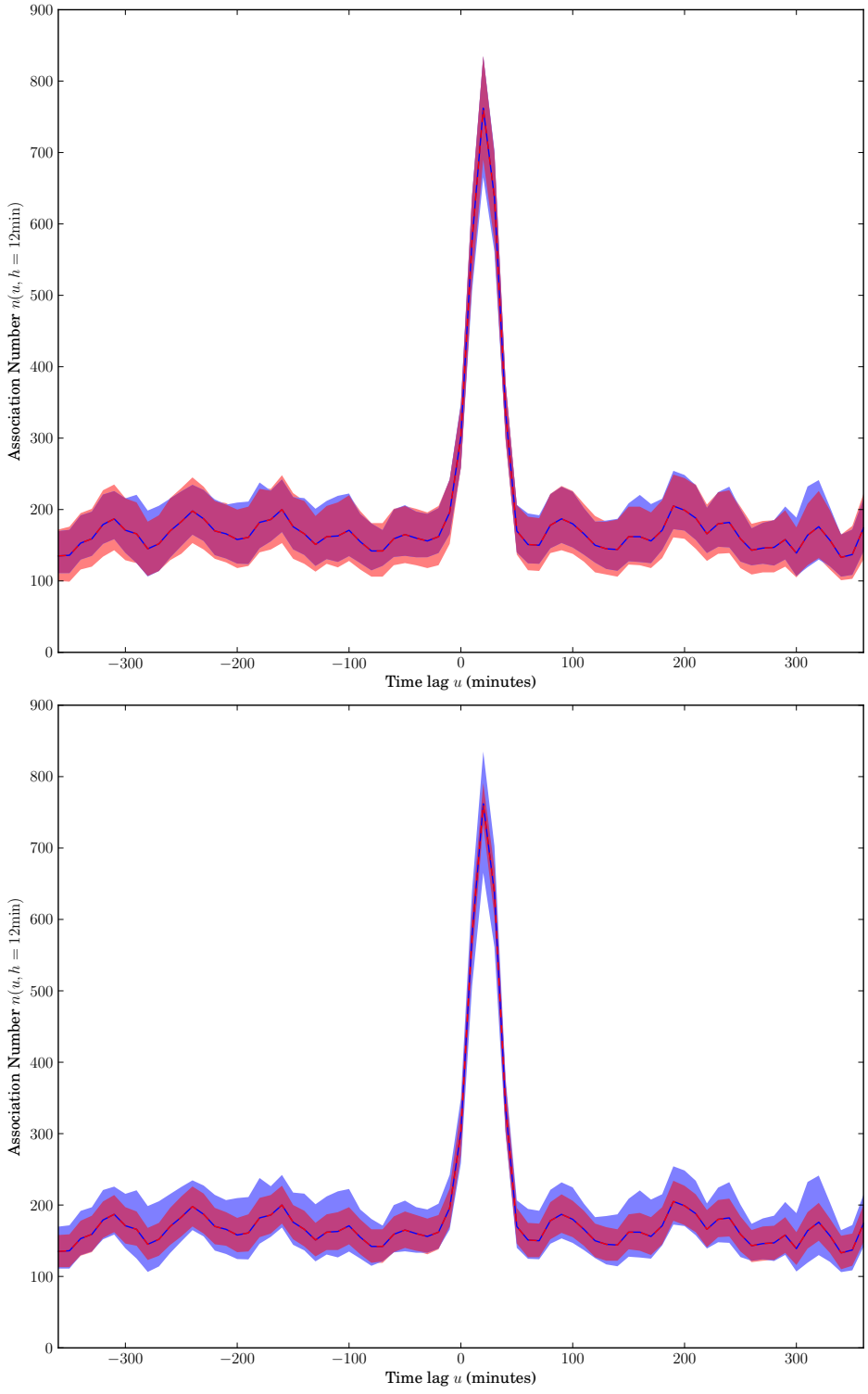

Figure 4: MBB 95% confidence intervals, in blue, compared to bootstrapping on the association number, in red; overlap is in magneta. Top is in the "forward" sense, bottom in the "backward" sense.

Table 1: Comparison between association number 95% c.i. and asymptotic or expected values. $n(u \to \infty)$ denotes comparison to the asymptotic association number (section 2.1) and $n_{expected}$ to the expected association number (equation 4). Restricted to high absolute lags (120 min. or greater), where random variation should yield the theoretical percentages within, below, and above the c.i. Confidence intervals calculated from surrogates of the series A individual associations ($n*_A$), series B individual associations ($n*_B$), or blocks generated by the moving block bootstrap (n_{MBB}). Ranges given are 95% confidence interval for this percentage; see text

	c.i. method	within c.i.	below c.i.	above c.i.
$n(u \to \infty)$	$n*_A$	98% (94–100)	2% (0–6)	0%
$n(u \to \infty)$	$n*_B$	82% (70–92)	10% (2–18)	8% (2–16)
$n(u \to \infty)$	$n*_{MBB}$	92% (84–98)	8% (2–16)	0%
$n_{expected}$	$n*_A$	100%	0%	0%
$n_{expected}$	$n*_B$	82% (70–92)	8% (2–16)	10% (2–18)
$n_{expected}$	$n*_{MBB}$	92% (84–98)	6% (0–14)	2% (0–6)
	Theoretical	95%	2.5%	2.5%

An additional check based on the definition of the confidence interval can be made: in the asymptotic region, where variability is expected to arise from chance, the asymptotic association number ($n(u \to \infty)$, section 2.1) should fall within the 95% confidence interval about 95% of the time, below it about 2.5% of the time, and above it about 2.5% of the time. For the 50 lags in the asymptotic region (absolute lags >= 120 min.), table 1 shows what fraction of the association numbers fell within the 95% confidence interval, calculated from surrogates of the series A individual associations ($n*_A$), series B individual associations ($n*_B$), or blocks generated by the moving block bootstrap (n_{MBB}). A similar comparison is made for the expected association number at each lag ($n_{expected}$, equation 4). The ranges given are 95% confidence intervals, calculated by bootstrapping 40000 surrogates from the asymptotic lags and finding the fractions for each surrogate. This process gives, by definition, no knowledge of uncertainty for the 100% or 0% case. These figures show that the MBB asymmetry arises from a bias in this implementation, not a true asymmetric confidence interval. Bootstrapping on the individual associations shows no such bias, but bootstrapping over series B yields confidence intervals which exclude the asymptotic (or expected) number significantly more frequently than 95%. We seek an explanation why bootstrapping on series A should be more trustworthy.

The essential question for the applicability of the standard bootstrap is the independence of samples, i.e. does the individual association $c_{A,i}$ depend on the previous association $c_{A,i-1}$? If A-type events are sufficiently separated in time, then no B event would be associated with more than one A event, breaking any linkage between the two. Indeed, the series were constructed to have wide spacing on A: on average 120 hours between events compared to the analysis window of 12 min. By contrast, the B events were selected to cluster strongly around each A event. Thus the individual associations for B events are not independent of each other, whereas the individual associations for A events are independent. By the properties of the bootstrap, then, bootstrapping on the independent A individual associations should be more reliable, a choice confirmed by the above analysis.

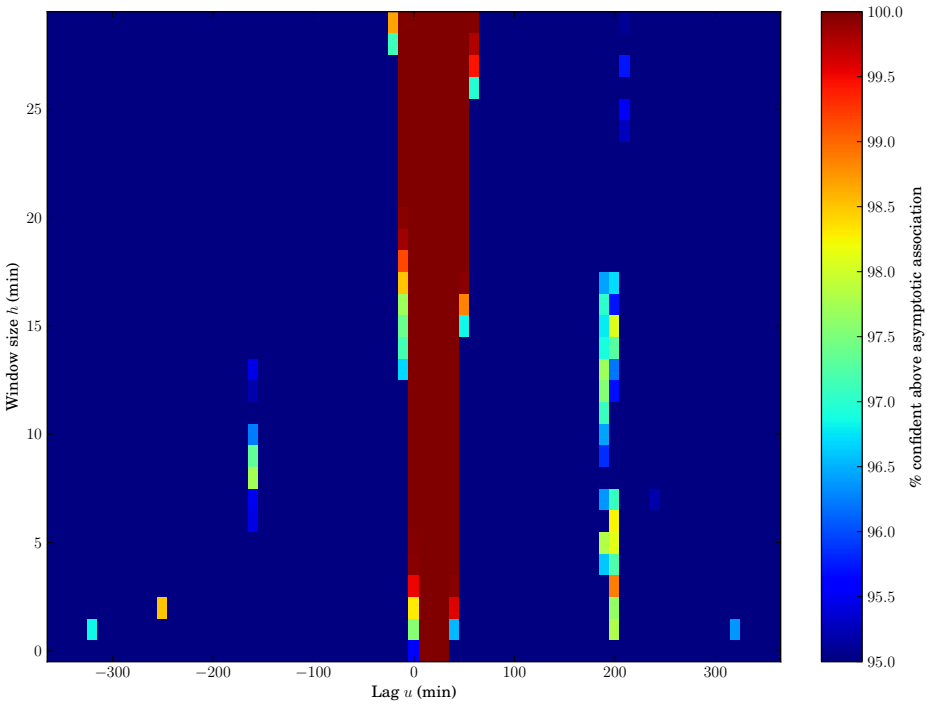

Figure 5: Confidence that the actual association number is above the asymptotic association, as a function of lag and window size.

4 Selection of window size

The choice of window is somewhat arbitrary, although it can be motivated by knowledge of uncertainties in timing or variability in the response of the underlying system. An overly narrow window will exclude real coincidences but provide sharper peaks in the association number as a function of lag; an overly broad window will yield uncertain estimates of the most significant lag. *Hsu and McPherron* [2002] presented a graphical method of searching for the optimal window size. They displayed the probability of the association number being above the asymptotic number as a function of both lag and window size, based on Poisson statistics. This approach can be adapted to the bootstrap method, with the results shown in figure 5. The peak at 20 min. lag is very apparent, and it is narrower for small window sizes. The bootstrap requires the sample be representative of the population, and there are few samples at the extrema of the distribution, so it cannot determine significances close to 100%. The method (and the plot) saturates as a result. Figure 6 illustrates one approach to finer-grained optimization of window size. Here the ratio of the calculated association number to the upper 95% confidence bound is plotted, and the sharper peak at small window sizes is well recovered. The method of figure 5 provides a first-cut minimum significance test, and figure 6 provides a means of maximizing the peak for windows meeting this minimum test.

Just as knowledge of the system motivates the initial choice of window, the optimal window size can be interpreted to infer some underlying information. It may be indicative of a range of delays in the process being measured; the width of the association peak is also useful for this information.

It should be noted that if the data are used to define the best size of search window, then the

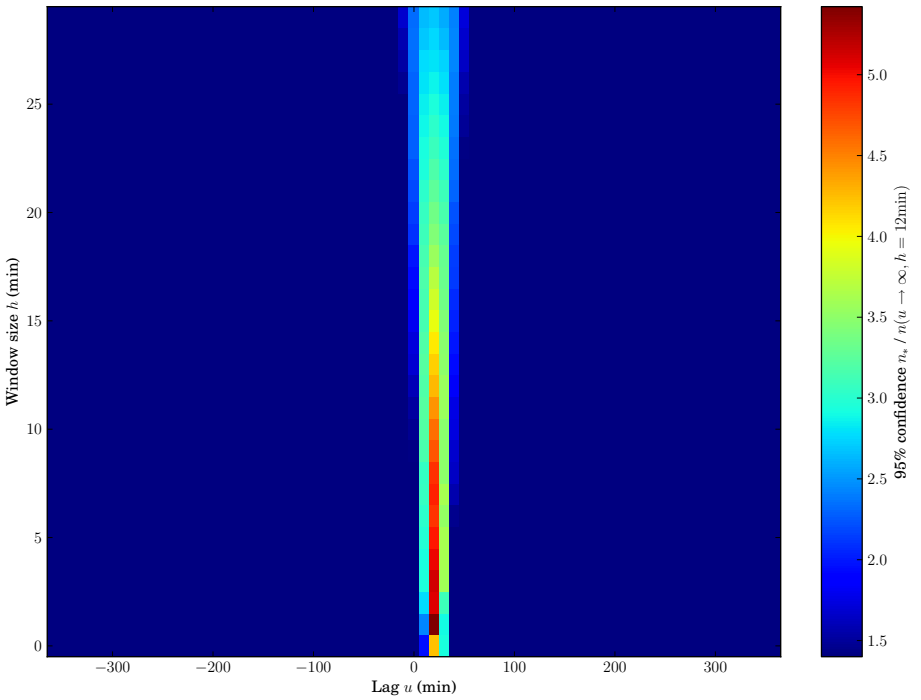

Figure 6: Ratio of the calculated association number's 95% confidence interval lower limit to the asymptotic association number, as a function of lag and window size.

testing of the hypothesis should be performed with a different, independent data set; defining the hypothesis test using the data from which the hypothesis was formulated artificially increases the likelihood of finding a significant result.

5 Future refinements

There are several potential enhancements to the techniques presented in this report, which we intend to add to SpacePy.

- Bootstrapping on the individual associations is fast, simple, and well-justified in certain cases, but a robust MBB implementation would allow reliable association analysis in the case of highly clustered events. In particular, a better implementation would eliminate the biased confidence intervals shown in figure 4.

- Periodic or quasi-periodic signatures often appear in the association number as a function of lag, extending into the asymptotic region. Initial attempts to relate these periodicities to periodicities on the input yielded mixed results. Properly filtering these periodicities may allow the confident recovery of weaker associations.

- Speeding up the bootstrap would open up more areas of study. As the association and bootstrapping code in SpacePy has evolved, we have implemented significant speedups, but clever algorithms may improve matters further. On modern machines with small enough data sets, the possibility exists to calculate all possible surrogates [*Diaconis and Holmes*, 1994a],

providing the best-possible bootstrap estimates for a particular sample. For larger samples, it may be possible to provide better coverage of the surrogate space with nonrandom selection of properly "spaced" surrogates [*Diaconis et al.*, 1996], similar to using Sobol sequences for other Monte Carlo problems. Where a high level of certainty is required for a confidence interval at the expense of the rest of the distribution, *Diaconis and Holmes* [1994b] presented a method to focus bootstrap computations on the tails.

6 Summary

This report presents techniques for determining potential associations between two types of event, a time lag in that association, and a confidence that such association is beyond that expected from chance. They are useful in a wide variety of applications. In the sciences, they can be used to find connections and potential causality between natural processes. They could also be used to evaluate the effectiveness of intentional action: is an advertising campaign associated with shoppers visiting a store? Is there a negative association (decrease in the association number) between writing speeding tickets and instances of speeding drivers? For how long does this effect hold?

We examined several approaches for determining confidence intervals on the association number. Bootstrapping on the individual associations for one of the series provides reliable confidence intervals, provided that the individual associations are independent of each other. This requirement is fulfilled if that series has events widely separated in time compared to the window half-width.

These powerful tools come with some caveats. As with all statistical procedures, they require sufficient sampling to be trustworthy. Appropriate confidence intervals are a guide, but the bootstrap itself is not necessarily reliable for very small samples. The elements of the series to which the bootstrap is applied must be independent, unless a technique like the MBB is properly applied. Applying any statistical tool indiscriminately in a hunt for significance will likely mislead: performing 20 statistical tests at 95% confidence will normally discover a "significant" result from simple chance. Finally, we include the standard caveat that association neither proves nor implies causation: some third process may drive both of the processes under investigation.

A Code for the figures

Figures 2, 3, 5, and 6 were produced using the Point Processes in Python (PoPPy) module of SpacePy. (Figure 4 used experimental Moving Block Bootstrap code, which has not been incorporated.) The code in this appendix shows how to reproduce these plots. In the interest of readability, cosmetic options (such as setting axis labels) have been eliminated, as have commands to save the plots. This code requires SpacePy, NumPy, and matplotlib.

```
import bisect
import math

from spacepy import poppy
import matplotlib
import matplotlib.pyplot as plt
import numpy

#Create the series
```

```
gauss = lambda x: math.exp(-(float(x - 20) ** 2) / (2 * 10 ** 2)) / \
        (10 * math.sqrt(2 * math.pi))
lags = range(-60 * 6, 60 * 6 + 1, 10) #minutes, up to quarter day
numpy.random.seed(1337)
series1 = numpy.random.randint(60 * -15 * 6, 60 * 105 * 6 + 1,
    [60 * 6])
series1.sort()
series1 = numpy.array(series1, dtype='float64')
randvals = numpy.random.rand(60 * 90 * 6 + 1)
series2 = numpy.fromiter((
    i for i in xrange(0, 60 * 90 * 6 + 1)
    if gauss(i - series1[bisect.bisect_right(series1, i) - 1]) +
       gauss(i - series1[bisect.bisect_left(series1, i)])
       > 0.25 * randvals[i]
    ), numpy.float64, count=-1)

#Create a PPro object from the series
pop = poppy.PPro(series1, series2, lags=lags, winhalf=12.0)
pop.assoc() #Perform association analysis

#Figure 2
pop.plot(norm=False)

pop.aa_ci(95, n_boots=4000) #Generate confidence intervals

#Do the same for the series in the reverse order
poprev = poppy.PPro(series2, series1, lags=lags[-1::-1], winhalf=12.0)
poprev.assoc()
poprev.aa_ci(95, n_boots=4000)
poprev.lags = lags

#Figure 3
poppy.plot_two_ppro(pop, poprev, ratio=1.0)
plt.show()

#Set up for window search algorithm
pop_2d = poppy.PPro(series1, series2, lags=lags, winhalf=12.0)
windows = numpy.array(range(30), dtype='float64')
(low, high, percentile) = pop_2d.assoc_mult(windows, n_boots=10000,
                          seed=1337)

#Figure 5
pop_2d.plot_mult(windows, percentile, min=95.0)
plt.show()

#Figure asymptotic association for all windows
```

```
asymptotes = numpy.empty([30], dtype='float64')
for i in range(len(windows)):
    junk = poppy.PPro(series1, series2, lags=lags, winhalf=windows[i])
    junk.assoc(h=windows[i])
    asymptotes[i] = junk.asympt_assoc

#Ratio of bottom of c.i. to asymptote
ratios = numpy.empty(low.shape, dtype='float64')
for i in range(len(windows)):
    ratios[i, :] = low[i, :] / asymptotes[i]

#Figure 6
pop_2d.plot_mult(windows, ratios, min=1.4,
                 cbar_label=r'95 percent confidence / asymptote')
plt.show()
```

B The bootstrap function in SpacePy

The `aa_ci` method of `PPro` objects selects surrogates from the individual associations and sums each surrogate series to find association numbers based on the surrogates, determining the 95% confidence interval from the 2.5^{th} and 97.5^{th} percentile of the surrogate association numbers. The underlying function, `boots_ci`, is available for more general use. It requires a sequence from which to select surrogates, the number of surrogate series to produce, a desired confidence interval, and any Python function which can be applied to a series for the desired metric. The following code implements the example of section 2.2.

```
import numpy
from spacepy import poppy

seq = [1.0, 2.0, 5.0, 5.0, 6.0]
#numpy.mean and numpy.median find mean and media of a sequence
print(numpy.mean(seq))
print(numpy.median(seq))

#Find 95 percent confidence intervals on mean, median
#Select 10000 surrogates
(lo, hi) = poppy.boots_ci(seq, 10000, 95, numpy.mean)
print('c.i. ' + str(lo) + ' through ' + str(hi))
print(poppy.boots_ci(seq, 10000, 95, numpy.median))

#What is the probability that the mean is above 4?
(lo, hi, prob) = poppy.boots_ci(seq, 10000, 95, numpy.mean, target=4.0)
print(str(prob) + ' percent chance mean greater than 4.')
```

The Python interface is backed by a fast C-based implementation, which takes full advantage of modern multicore and multiprocessor machines.

References

Brillinger, D. R. (1976), Measuring the association of point processes: a case history. *The American Mathematical Monthly*, *83*(1), 16–22.

Cox, D. R. (1955), Some statistical methods connected with series of events. *Journal of the Royal Statistical Society, Series B*, *17*(2), 129–164.

Diaconis, P. and B. Efron (1983), Computer-intensive methods in statistics. *Scientific American*, *248*(5), 116–130, doi:10.1038/scientificamerican0583-116.

Diaconis, P. and S. Holmes (1994a), Gray codes for randomization procedures. *Statistics and Computing*, *4*(4), 287–302, doi:10.1007/BF00156752.

Diaconis, P. and S. Holmes (1994b), Three Examples of the Markov Chain Monte Carlo Method. In Aldous, D., P. Diaconis, J. Spencer, and J. M. Steele, eds., *Discrete Probability and Algorithms*, 43–56, Springer Verlag, New York.

Diaconis, P., S. Holmes, S. Janson, S. P. Lalley, and R. Pemantle (1996), Metrics on compositions and coincidences among renewal sequences. In Aldous, D. and R. Pemantle, eds., *Random Discrete Structures*, 81–102, IMA publications, Springer Verlag, New York.

Efron, B. (1979), Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26.

Hsu, T. and R. L. McPherron (2002), An evaluation of the statistical significance of the association between northward turnings of the interplanetary magnetic field and substorm expansion onsets. *Journal of Geophysical Research (Space Physics)*, *107*, 1398, doi:10.1029/2000JA000125.

Künsch, H. R. (1989), The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, *17*(3), 1217–1241, doi:10.1214/aos/1176347265.

Lahiri, S. N. (2003), *Resampling methods for dependent data*. Springer-Verlag, New York.

Liu, R. Y. and K. Singh (1992), Moving blocks jackknife and bootstrap capture weak dependence. In Lepage, R. and L. Billard, eds., *Exploring the limits of the bootstrap*, 225–248, Wiley, New York.

Morley, S. K. and M. P. Freeman (2007), On the association between northward turnings of the interplanetary magnetic field and substorm onsets. Geophys. Res. Lett., *34*, L08104, doi:10.1029/2006GL028891.

Morley, S. K., J. Koller, D. T. Welling, B. A. Larsen, M. G. Henderson, and J. T. Niehof (2011), SpacePy - A Python-based library of tools for the space sciences. In *Proceedings of the 9th Python in science conference (SciPy 2010)*, Austin, TX.

Singh, K. (1981), On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics*, *9*(6), 1187–1185, doi:10.1214/aos/1176345636.


Los Alamos
NATIONAL LABORATORY
EST. 1943