ECRI Release

THE LEAD PROFILE AND OTHER NON-PARAMETRIC TOOLS TO EVALUATE SURVEY SERIES AS LEADING INDICATORS

Anirvan Banerji
Economic Cycle Research Institute
New York

24th CIRET Conference
Wellington, New Zealand
March 17-20, 1999

TESTING FOR CYCLICAL LEADS

The hallmark of a cyclical leading indicator is the property that its cyclical turning points lead cyclical turning points in the economy. However, there are no well-known methods to test whether these leads are statistically significant.

It has long been recognized that leading indicators can be a valuable forecasting tool for forecasting cyclical turning points. However, they have not always been properly evaluated. One method of evaluating leading indicators that has gained some popularity in recent years is the Granger causality test. It is thus interesting to note what Granger and Newbold (1986) have to say about the difficulty of evaluating the index of leading indicators (LEI):

> The index of leading indicators has become a widely quoted and generally trusted forecasting tool. However, it has been rather misinterpreted. The index is intended only to forecast the timing of turning points and not the size of the forthcoming downswing or upswing nor to be a general indicator of the economy at times other than near turning points. Because of this, evaluation of the (LEI) by standard statistical techniques is not easy.

This difficulty in evaluation has often led to flawed assessments of the performance of leading indicators, not necessarily based on their ability to anticipate turning points. Part of the problem has been a lack of familiarity with the standard methods of identifying turning points.

Yet, since leading indicators are meant primarily to forecast business cycle turning points, the identification of turning points in time series is a *sine qua non* for an appropriate evaluation of their forecasting performance. In fact, an objective algorithm for turning point identification, based on a systematic codification of the judgmental procedures used for decades at the NBER, was devised almost three decades ago (Bry and Boschan, 1971), shortly after the creation of the LEI. The Bry-Boschan procedure has certainly stood the test of time.

Geoffrey Moore, who helped create the LEI (Moore and Shiskin, 1967), used the Bry-Boschan procedure extensively in the decades following its creation (e.g., Klein and Moore, 1985). Other users have included King and Plosser (1989), who provide a description of the procedure. Watson (1994) points out that the Bry-Boschan procedure provides a good way to define turning points since it is based on objective criteria for determining cyclical peaks and troughs.

The objective (though not mathematically simple) definition of turning points given by Bry and Boschan's algorithmic formulation of the classical NBER procedure makes it possible to evaluate the performance of leading indicators in terms of an objective measure of the leads of leading indicators at turning points. In that sense, the Bry-Boschan procedure permits a more appropriate evaluation of the performance of leading indicators.

Given the cyclical turning points of a potential leading indicator, it is possible to measure the lead of that indicator at each business cycle turning point. However, many leading indicators cover only a small number of cycles. Thus the evaluation of leading indicators by parametric statistical methods is usually constrained by the limited number of cyclical turning points covered by the data. In addition, the need to make a heroic assumption that the probability distribution of the leads has a standard functional form also precludes the use of parametric tests of statistical significance.

This paper suggests simple nonparametric tests to evaluate the cyclical leads of leading indicators, and introduces lead profile charts that graphically depict these leads in probabilistic terms, to aid in the selection and evaluation of leading indicators.

THE PROBLEM

A number of considerations go into the evaluation of any time series as a cyclical leading indicator. There are at least two criteria on which survey series usually excel: One, unlike many other potential leading indicators, they are typically never revised; and two, they are very promptly available.

However, survey series must still be evaluated on the magnitude and variability of their leads compared with a reference cycle (such as the business cycle) at cyclical turns, as well as their leads compared with one another when two or more survey series are being compared. In all of these cases, the magnitude (and even the direction) of the lead may vary from one turn to the next. The issue, then, is the statistical significance of the leads, or of the difference in leads, as the case may be. Another major issue is the difference in the variability of the leads of two potential leading indicators that may have roughly the same magnitude of lead.

We have cited Granger and Newbold (1986) who suggest, in effect, that standard statistical approaches to the evaluation of leading indicators may be fraught with problems. The simpler classical approach of just measuring the mean and standard deviation of the leads does not result in tests of statistical significance without an assumption that the probability distribution of the leads has a standard functional form. Thus, no tests of significance can usually be performed. Under such circumstances, simple nonparametric tests may be the most appropriate solution.

APPROPRIATE NONPARAMETRIC TESTS

Nonparametric tests are often called "distribution-free" because they do not assume that the observations were drawn from a population distributed in a certain way, e.g., from a normally distributed population. These tests also do not require the large samples needed to reliably estimate parameters of distributions assumed in parametric tests. Such tests should therefore be uniquely suited to testing the significance of leads, which may be small in number, and for which the probability distribution function is quite unknown.

Since the leads in question are differences in timing at cyclical turns (between a pair of indicators, for example), the appropriate nonparametric tests are those applicable to matched pairs of samples. The most powerful tests in this class assume interval scaled data (like temperature in degrees Celsius) where equal intervals at any point in the scale imply equal differences. Leads measured in months or quarters are at least interval scaled, so such tests can be used with data on leads.

The most appropriate test to assess the significance of leads within this class is the Randomization test for matched pairs. This test has a power-efficiency of 100%, because it uses all the information in the sample (Siegel, 1956), but it does not lend itself to manual computation for sample sizes greater than about nine pairs. (For an example of a manual application and computational considerations see Appendix). In such cases, a simple computer program can be used.

One alternative is the Wilcoxon matched-pairs signed-ranks test (Wilcoxon, 1945), which assumes that the data are at least ordinal, has a power-efficiency of 95% (Mood, 1954), and can be performed manually up to sample sizes of 25 pairs.

There is a much smaller choice of tests for the difference in the variability of leads. One appropriate test is the Moses test, which is based on the pooled ranks of the leads of the two leading indicators being compared.

THE RANDOMIZATION TEST FOR MATCHED PAIRS

The Randomization test (Fisher, 1935) is a simple and elegant way to test the significance of leads. It is easy to use manually for sample sizes under 10, and a simple computer program can be used for larger sample sizes, so a description of the procedure and rationale should be instructive.

The first step is to calculate the difference in timing at turns, that is, the leads of one indicator over another, or over the business cycle turning points. The null hypothesis, that these differences are not statistically significant, is to be tested against the alternative hypothesis that the leads are significant.

Now, some of the differences calculated in the first step may be positive, others negative. If the null hypothesis is true, the positive differences are just as likely to have been negative, and vice versa. So if there are N differences (from N pairs of observations), each difference is as likely to be positive as negative. Thus, the observed set of differences would be just one of 2^N equally likely outcomes under the null hypothesis.

Also, under the null hypothesis, the sum of the positive differences would, on average, equal the sum of the negative differences, so the expected sum of the positive and negative differences would be zero. If the alternative hypothesis were true, and the leads were positive and significant, the sum would very likely be positive.

The second step, therefore, is to sum the differences, assigning positive signs to each difference; then to switch the signs systematically, one by one, to generate all the outcomes which result in sums as high or higher than that observed. If there are R such outcomes, then the probability of the observed outcome (or a more extreme outcome) under the null hypothesis is $(R/2^N)$. In other words, the null hypothesis can be rejected at the $100(1-(R/2^N))$% confidence level.

LEAD PROFILES

So far, the discussion has focused on the confidence level at which the null hypothesis ("leads not significantly different from zero") can be rejected in favor of the alternative hypothesis ("leads significantly greater than zero months"). Now, even if it is established that the leads are significantly greater than zero months, it might be interesting to know how much greater than zero months the leads are likely to be – for example, whether the leads are also significantly greater than one month.

This is easy to determine. All one needs to do is to subtract one month from each of the differences in timing at turns (already calculated in the first step of the Randomization test). Then, as before, one finds the confidence level at which the null hypothesis is rejected in favor of the alternative hypothesis that the difference in timing at turns significantly exceeds one month.

In this way one can also determine the confidence levels for the hypotheses that the leads exceed 2,3,4, …. K months – simply by subtracting 2,3,4, …. K respectively from the original differences before performing the Randomization test. We call this full set of confidence levels a "lead profile".

Ordinarily, it would be rather tedious to calculate the lead profile for each indicator against business cycle turning points or against another indicator. But this happens to be the kind of computation ideally suited to a computer program, and our program typically computes a complete lead profile in a few seconds.

The lead profile can also be graphically represented in bar charts or "lead profile charts". Figure 1 is the lead profile chart for the lead of the University of Michigan's Consumer Expectations Index compared with the Conference Board's version at business cycle troughs. The question answered by this chart is whether the difference between the leads of the two indexes is statistically significant. The lead profile is a graphical depiction of the leads in strictly probabilistic terms, that aids meaningful comparisons between the indexes. It is clear from this chart that the difference between the leads is not statistically significant, since the confidence level that the Michigan index leads the Conference Board Index is only 20%.

Figure 2 depicts the lead profile of the National Association of Purchasing Management (NAPM) Index compared with U.S. business cycle turns. It shows that there is a 95% confidence level that the lead of this index at business cycle turns is at least seven months.

Similarly, Figure 3 shows the lead profile of the NAPM New Orders Diffusion Index against U.S. business cycle turns. It also shows that there is a 95% confidence level that the lead of this index at business cycle turns is at least seven months.

Figure 4 depicts the lead profile of the NAPM Production Diffusion Index against U.S. business cycle turns. It shows that there is a 95% confidence level that the lead of this index at business cycle turns is at least six months.

Figure 5 depicts the lead profile of the NAPM Inventory Diffusion Index against U.S. business cycle turns. It shows that there is a 95% confidence level that the lead of this index at business cycle turns is at least three months.

While these lead profiles show primarily the confidence level in different durations of lead, the rapidity of the decline also casts some light on the variability of the lead. For example, the lead profile of the overall NAPM Index (Figure 2) shows a drop of only 46% between six and 12 months. The corresponding number is 67% for the NAPM Production Diffusion Index (Figure 4), 71% for the NAPM Inventory Diffusion Index (Figure 5), and 80% for the NAPM New Orders Diffusion Index (Figure 3). In that sense, the leads of the latter indexes are clustered more closely than they are for the overall index.

However, the lead profile does not provide a formal statistical test of the difference in the variability of the leads of two indexes. An appropriate nonparametric test for that purpose is the Moses test.

THE MOSES TEST

The Moses test (Moses, 1952) is designed for a situation when the there may not be any significant difference in central tendency between two samples (e.g., the leads of two competing leading indicators), but the leads of one may be more spread out or variable than the other. Thus, even if the lead profile does not reveal any significant difference in the leads of two cyclical leading indicators, the leads of one indicator may be less variable than the leads of the other.

The Moses test focuses on the span or the spread of the leads with which the comparison is being made. Thus, if there are n_C such leads, and there are n_E leads of the "experimental" leading indicator being compared to the reference series, the $n_E + n_C$ leads are arranged in order of increasing size. If the null hypothesis, that the leads of the two indicators come from the same population, is true, the E's and C's should be well mixed in the ordered series.

Under the null hypothesis, therefore, the very long leads, the very short leads, as well as the moderate leads, should all contain a mixture of E's and C's. Under the alternative hypothesis, most of the E's will be low, or most of the E's will be high, or most of the E's will be low or high with the C's concentrated in the middle. Under these three conditions, the C's will be concentrated at the high end, the low end, and the middle region, respectively. The Moses test determines whether the C's are so closely compacted or congested relative to the $n_E + n_C$ leads as to call for a rejection of the null hypothesis that both E's and C's come from the same population.

To perform the Moses test, the leads from the E and C indicators are combined and arranged in a single ordered series, retaining the identity of each lead as having come from the C indicator or the E indicator. Then the span of the C leads is determined by noting the lowest and highest C leads and counting the number of leads between them, including both extremes. Thus the span s' is defined as the smallest number of consecutive leads in an ordered series needed to include all the C leads. Since the sampling distribution of s' is known (Moses, 1952), it may be used for tests of significance.

Because s' is essentially the range of the C leads, and the range is known to be unstable, the modification suggested by Moses is to pick an arbitrary small number h, in advance of the analysis. A value such as 1 or 2 is typical. Then, the span s_h of the C leads is determined after dropping the h most extreme C ranks. Then,

$$g = s_h - (n_C - 2h) \quad \text{.......................................} \quad (1)$$

The probability under the null hypothesis of observing the observed value of s_h or less is:

$$p(s_h \leq n_C - 2h + g) = \frac{\sum_{i=0}^{g} \binom{i + n_C - 2h - 2}{i} \binom{n_E + 2h + 1 - i}{n_E - i}}{\binom{n_C + n_E}{n_C}} \quad \ldots\ldots\ldots \quad (2)$$

If p is smaller than or equal to the significance level, the null hypothesis is rejected in favor of the alternative hypothesis that the E leads are more spread out than the C leads.

As shown in part C of the Appendix, when the Moses test is conducted on the consumer expectations survey series produced by the University of Michigan and the Conference Board, respectively, it shows no significant difference in the spread of their leads at U.S. business cycle turning points. Specifically, the probability is 0.84 that the observed leads would have been observed given the null hypothesis that the spread of the leads of the two indicators is the same. In the case of a comparison between the NAPM Index and the NAPM New Orders Diffusion Index (calculations not shown), the comparable probability is less than 0.13. While this is not statistically significant, it is certainly very suggestive that the leads of the New Orders Index are less spread out than those of the overall index. Given that for both, the lead profiles show the confidence level for a lead of seven months (but no higher) to be above 95%, and both have a median lead of eight months at business cycle turns, the New Orders Index may be preferred as a leading indicator to the overall index. The Moses test allows such distinctions to be made on the basis of a formal statistical test.

FIGURE 1: Lead Profile: Lead of Univ. of Michigan Consumer Expectations Over Conference Board Consumer Expectations

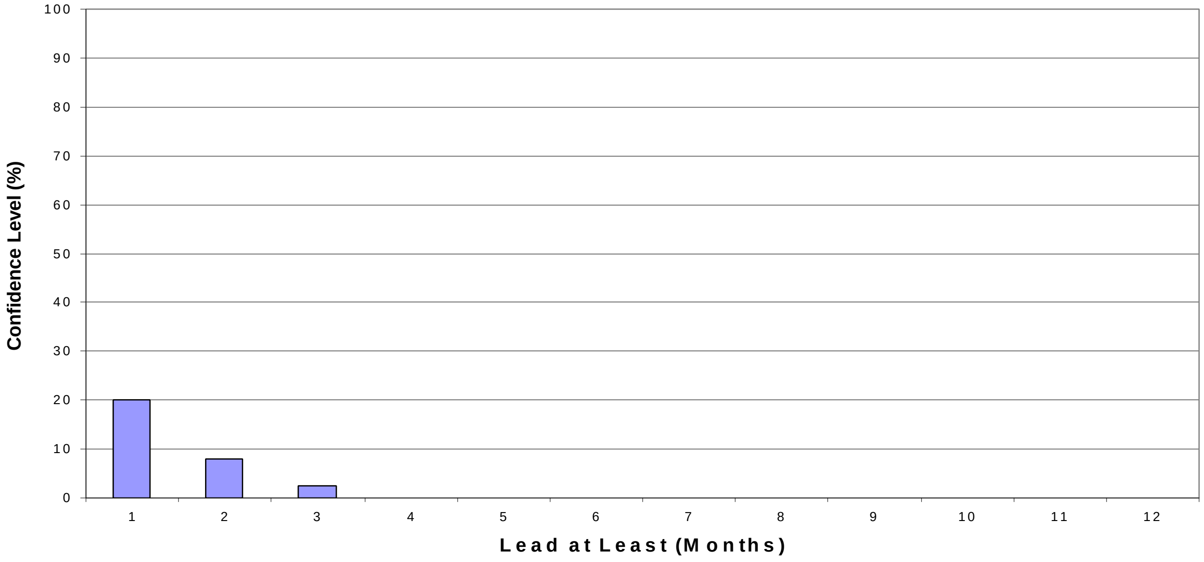


FIGURE 2: Lead Profile of NAPM Index: Lead Over U.S. Business Cycle Turns

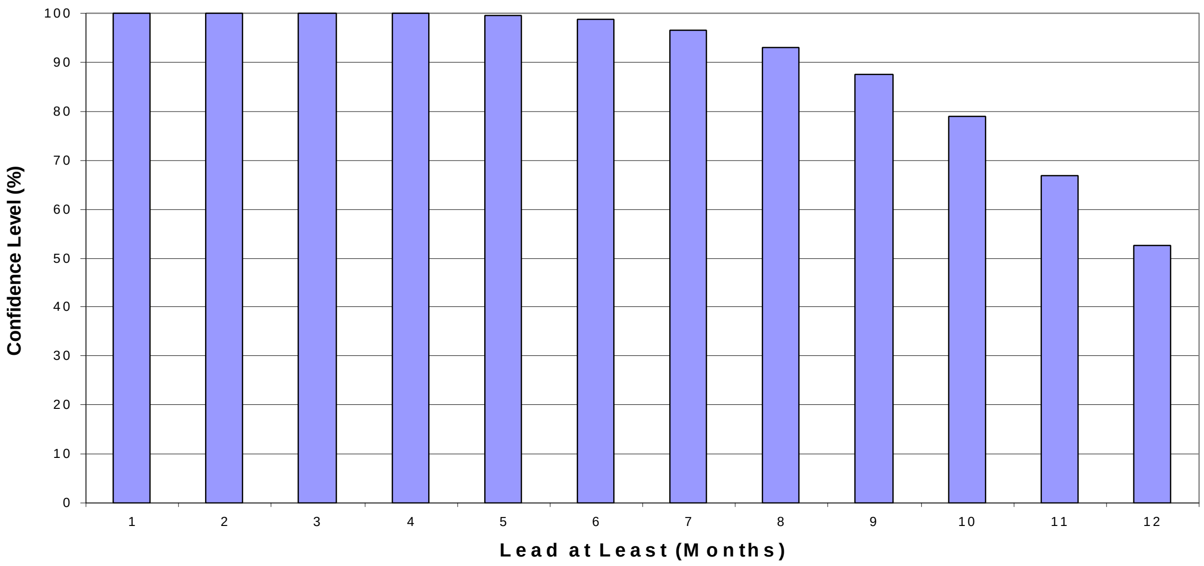


FIGURE 3: NAPM New Orders Diffusion Index: Lead Over U.S. Business Cycle Turns

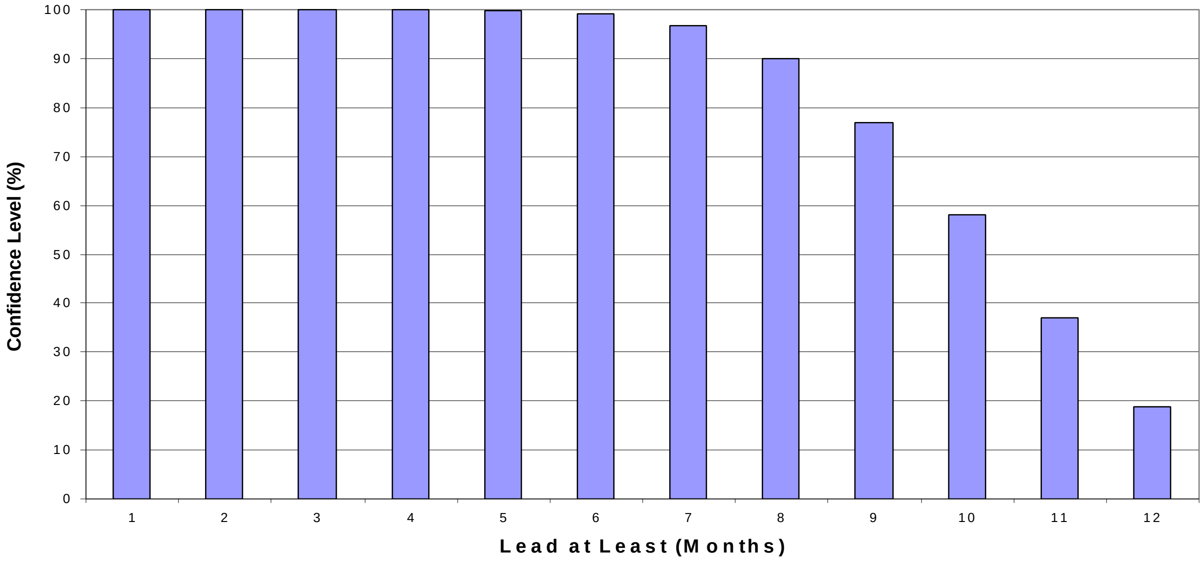


FIGURE 4: Lead Profile of NAPM Production Diffusion Index: Lead Over U.S. Business Cycle Turns

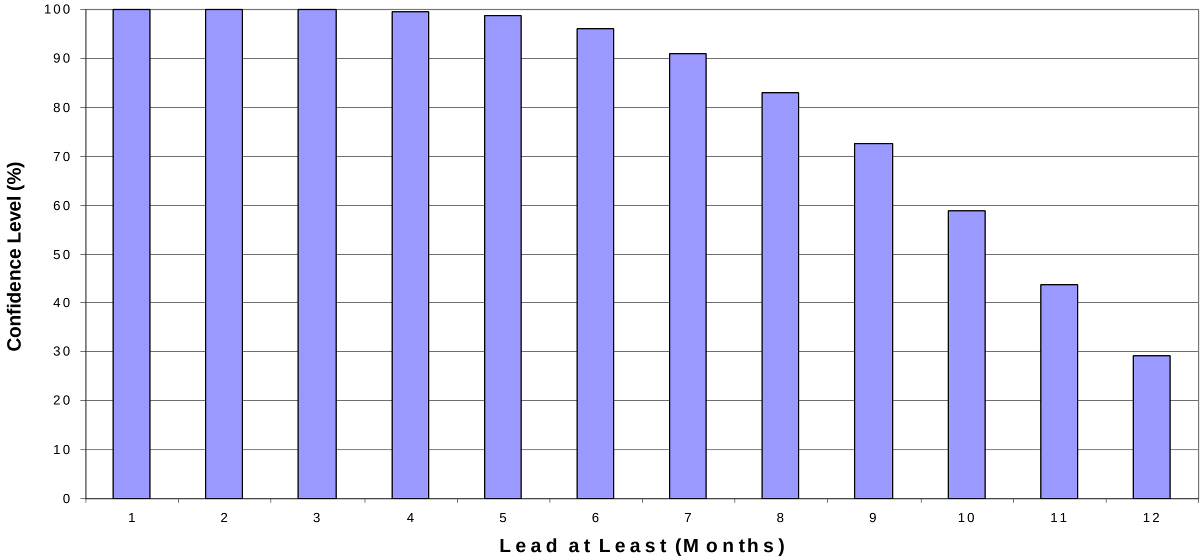
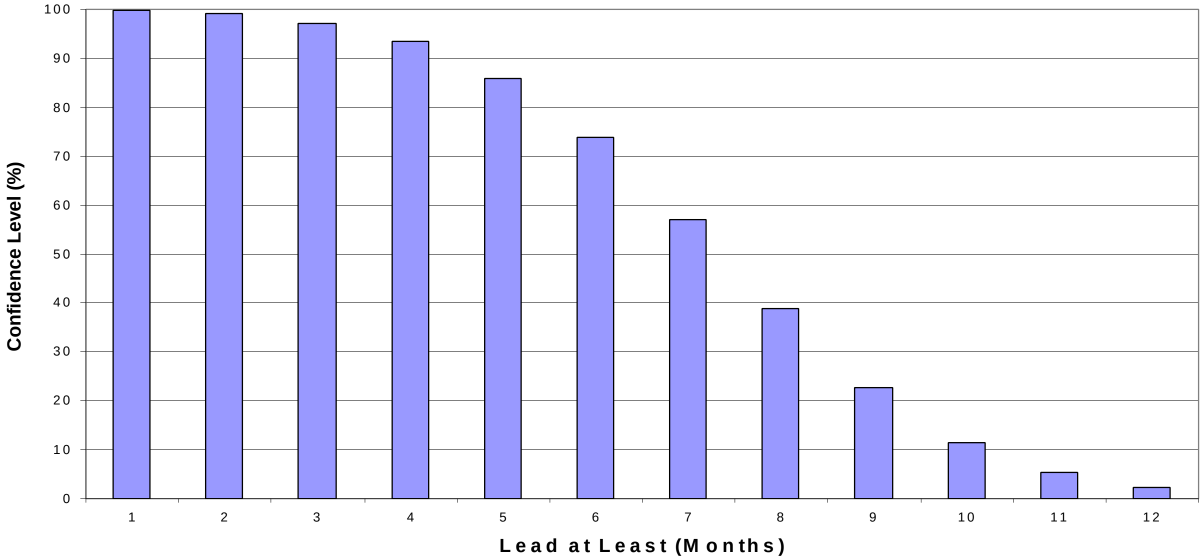


FIGURE 5: Lead Profile of NAPM Inventory Diffusion Index: Lead Over U.S. Business Cycle Turns



CONCLUSIONS

Lead profile charts use as input just the information on the length of the leads at each turning point. However, by gleaning statistical inferences from the data rather than relying solely on averages, and by displaying the results graphically, they afford additional insights into the significance of leads.

The lead profile concentrates on the magnitude of leads, and can test whether the lead of one time series is significantly greater than another's. It cannot, however, tell whether there is a significant difference in the variability of the leads of two time series that do not have significantly different leads. The Moses test is well suited for that purpose, and can therefore supplement lead profiles in evaluating cyclical indicators.

Lead profile charts do not directly address the issue of extra cycles or missed turns, as they start with just the matched pairs of observations. However, in the case of missed turns, the exclusion of potential data points results in a smaller sample size, and a consequent reduction in the confidence level, which is appropriately conservative.

The major advantage of lead profiles lies in the explicit statistical inferences that can be made about the significance of leads without making any assumptions about the probability distribution of leads, or any restrictions on sample size. These inferences can be made about the leads of a given cyclical indicator over a reference cycle, such as a set of business cycle turning points. They can also be made about the leads of one cyclical indicator over another, to assess whether one has significantly longer leads than the other. Moreover, it is convenient to put lead profiles in the form of bar charts, for easy and effective visual appraisal of the significance of lengths of leads.

The magnitude and variability of the leads are perhaps the two most important factors in the evaluation of survey series as leading indicators. Given the ease of using these simple yet powerful statistical tools for evaluating cyclical leads, there is a strong case for using them routinely whenever the performance of a leading indicator needs to be assessed.

APPENDIX

A. USING THE RANDOMIZATION TEST TO TEST THE SIGNIFICANCE OF LEADS OF A HYPOTHETICAL LEADING INDICATOR OVER BUSINESS CYCLE TROUGHS

The leads at troughs of this indicator compared to the business cycle troughs are 12, 4, 1, 0 and -27 months. The last figure represents a lag of 27 months. Although the convention is to use negative numbers for leads, and positive numbers for lags, it is simpler for the purpose of this exposition to think of leads as being positive, because we are, in general, concerned with the significance of leads, not lags.

The first step is to drop the zero-month lead from the analysis; keeping this observation would make no difference to the results, as is evident from the procedure for the Randomization test. Then N = 4, and the 4 observations are (12, 4, 1, -27), which add up to a sum of S = -10.

This sum S is now compared with the sums computed by starting with all positive numbers, and switching signs one by one so that the sums are in descending order until our sum of S = 10 is reached:

12	4	1	27	Sum = 44
12	4	-1	27	Sum = 42
12	-4	1	27	Sum = 36
12	-4	-1	27	Sum = 34
-12	4	1	27	Sum = 20
-12	4	-1	27	Sum = 18
-12	-4	1	27	Sum = 12
-12	-4	-1	27	Sum = 10
12	4	1	-27	Sum = -10 = S

Since R = 9 sums out of 24 (i.e., 16) possible combinations are greater than or equal to -10, the probability of such an outcome under the null hypothesis ("leads not significant") is 9/16 = 0.5625, so that the null hypothesis can be rejected only at the 100 (1-0.5625)% = 43.75% level of confidence. Hence, the null hypothesis is accepted for leads at troughs.

B. COMPUTATIONAL CONSIDERATIONS

The randomization test can be performed manually if the number of matched pairs is less than ten, as is often the case with turning points. For example, if the hypothesis needs to be tested at the 95% confidence level, the number of sums that need to be computed for nine matched pairs of observations is at most 5% of 2^9, that is 26, and possibly fewer. But the computational load increases with the lowering of the confidence level of the test, and increases exponentially with sample size. Also, the order in which signs are switched is important since the sums must be in descending order, and this complicates the procedure even further for large samples.

A computer program has therefore been prepared to calculate the confidence level at which the null hypothesis ("differences not significant") can be rejected, given a set of differences in the timing of turns of a pair of indicators.

If the sample size N is higher than 25 matched pairs, the normal approximation is appropriate, as long as the differences d_i are all about the same size, so that

$$(d_{max})^2/\Sigma(d_i)^2 = (5/2N)$$

where d_{max} is the largest observed difference. In that case, Σd_i is approximately normally distributed with mean zero and standard deviation $\sqrt{\Sigma(d_i)^2}$, so that

$$Z = (\Sigma d_i - \mu)/\sigma = \Sigma d_i/(\sqrt{\Sigma(d_i)^2})$$

is approximately normally distributed with zero mean and unit variance.

C. AN APPLICATION OF THE MOSES TEST

There are two popular measures of consumer expectations used in the U.S., one produced by the University of Michigan, and the other by the Conference Board. There is no significant difference between the magnitude of their leads against the business cycle, as shown by their lead profiles. Is there a significant difference between the variability of these leads?

The leads of the Michigan (m) index at business cycle turning points starting in 1969 are: 14, 9, 11, 5, 37, 2, 8, 8, 26 and 2 months. The corresponding leads of the Conference Board (c) index are: 10, 6, 15, 1, 38, 4, 2, 8, 18 and 5 months.

The m and c leads are now combined and arranged in a single ordered series, retaining the identity of each lead as having come from the m indicator or the c indicator:

38	37	26	18	15	14	11	10	9	8	8	8	6	5	5	4	2	2	2	1
c	m	m	c	c	m	m	c	m	m	m	c	c	m	c	c	m	m	c	c

Having chosen h=1 in advance, the span of ranks, after the most extreme-ranked m leads are dropped, comes to 15. The value of g, the amount by which this observed span exceeds n_m – 2h is (15 – (10 - 2)) = 7. Using these values in equation 2 yields a probability of 0.84 that the observed leads would have been observed given the null hypothesis that the spread of the leads of the two indicators is the same. Thus the null hypothesis is accepted.

