November 24, 2013

How Practical Is Statistical Significance?*

Ben Jacobsen Alexander Molchanov Cherry Zhang

Abstract

Statistical significance is a core concept in many sciences to distinguish between what is and what is not. But what are the practical implications of finding a significant effect, and is it more reliable when based on more observations or higher confidence levels? We consider forecast accuracy in a practical setting using significant monthly deviations from the mean in a more than 300 year long time series of stock market returns. While statistical significance in general implies a marginal improved probability of correct forecasts, results vary strongly depending on significance levels and number of observations used. Surprisingly, signals based on higher significance levels or more observations do not necessarily imply better forecasts.

* This version is for submission to the New Zealand Finance Colloquium. We would like to thank the participants of the Brown Bag seminar at Massey University (Palmerston North) for helpful comments and suggestions.

Ben Jacobsen, School of Economics and Finance, Massey University, Private Bag 102904, Auckland, New Zealand. E-mail: b.jacobsen@massey.ac.nz

Alexander Molchanov (corresponding author), School of Economics and Finance, Massey University, Private Bag 102904, Auckland, New Zealand. E-mail: a.e.molchanov@massey.ac.nz

Cherry Zhang, Nottingham University Business School, 199 Taikang East Road, Ningbo 315100, China. E-mail: cherry-yi.zhang@nottingham.edu.cn

1. Introduction

We tend to trust results more if their statistical significance is high and if these results are based on a large number of observations. And in theory we should. But in many natural experiments we cannot be sure about the exact data generating process and whether this process is constant over time. In situations like these, questions arise whether signals derived from more data are more accurate, whether higher significance levels imply better signals, and what role assumptions about the underlying data generating process may play.

Assuming that the return generating process would not change over time, we would expect we would expect higher forecast success rates for signals detected with higher confidence levels and longer measurement intervals. If the process changes, we would expect that there may be an optimal detection window length (and confidence level), as at some point the benefits of additional observations and accuracy are outweighed by changes in the return generating process.

While at first sight this may sound theoretical, it is a very practical question that any statistically trained investor building a financial forecasting system runs into. For instance, can an investor extract a valuable signal from the knowledge that over the last fifty years returns were higher in the month of December at a 5% significance level? Or is she better off considering signals based on ten years of observation, but at the 1% significance level?

In finance, predictability studies generally use expanding windows with a more or less arbitrary start dates. Different researchers use different model selection criteria and significance levels. This is all based on the assumption that the underlying return generating process has not changed, or that these changes can be explicitly modelled. However, as Timmerman (2007) shows, the process is constantly changing perhaps partially because of the forecaster's own efforts, and the best we can hope for is local predictability. The question then becomes what window length and significance level should we consider to optimally profit from any local predictability. To answer that question we consider one of the most basic statistical tests one can perform: testing for difference in mean.

We consider a long time series of stock market returns of over three centuries of monthly data. We first test whether a month shows a significantly different mean return over some measurement period. We then test whether that knowledge would lead to improved forecast for that month in the next year.

Whether some months are anomalous and show somehow significantly different returns from others has been one of the oldest questions in academic finance research. Wachtel (1942) first documented a January effect (average returns being higher in January than during the remainder of the year). But we are not interested in whether monthly anomalies exist as such or not (Zhang and Jacobsen (2012) show that this strongly depends on the time interval used to measure these deviations). Our interest lies in a more general question: What are the practical implications of finding some significant deviation over some measurement interval based on commonly used statistical models to describe data generating process, which in our case happens to be the return generating process in financial markets?

Zhang and Jacobsen (2012) find that even with more over three centuries of data it is hard to determine whether there monthly seasonal anomalies exist or whether these anomalies are simply in the eye of the beholder and seems to strongly depend on samples and sample sizes. Some months show strong significant deviations in some periods but not in others. Others months show a significant seasonal effect over the full three hundred years but still go for a century without any significant effect. However, they leave an interesting question unresolved. Given that these monthly anomalies may or may not exist what should investors do?

Trade on some of these anomalies once they pop up and if so is there an optimal sample size and or confidence level to be used to estimate whether an anomaly exists or not? If the existence of these anomalies is uncertain could an investor be better of trading on these effects or not. If the beholder is an investor what should she do? If there is a systematically significant effect over sample period of certain size is it worth trading on the effect? We address this question using the same long dataset as Zhang and Jacobsen (2012), because of its length seems well suited to address this question. It has seen major changes over the sample period which could have structural breaks, regime shifts, gradual changes or in the end no change at all. To illustrate: to get to the start or our series we have to look back past the September 11 attacks, past Margaret Thatcher, the hippies, the cold war, WWII and WWI, The Russian Revolution, Einstein's general relativity theory, the brothers Wright first flight.

We find that forecasting ability of significant monthly deviations depends on the length of the estimation window, but the pattern is far from clear. It appears that forecasting ability is the highest at relatively short (around 30 years) and extremely long (more than 250 years)

detection windows. However, the number of significant deviations for extremely long windows is small, and forecasting ability behaves rather erratically. Therefore, there is evidence that the optimal trade-off between accuracy in estimation (longer detection windows) and the fact that return generating process is likely changing appears to happen at relatively short detection window lengths.

Surprisingly, we document that deviations detected with higher significance levels *do not* necessarily result in higher forecasting success. Although forecasting accuracy of signals detected at 1% level is, on average, the highest, it is highly dependent on the detection window length. We detect significant deviations using OLS regressions with White standard errors. We document that using more complex techniques for signal detection (i.e., GARCH and Probit) does not, in general, improve forecasting accuracy.

The rest of the paper is organized as follows. Section 2 presents the data. Section 3 describes the methodology. We present our main findings in section 4. Section 5 describes robustness checks. Section 6 concludes.

2. Data

UK stock return index

For our experiment we rely on the 317-year index of monthly UK stock prices compiled by Global Financial Data from several different sources. Starting from 1693, the index basically covers the entire trading history of the UK equity market. This index is extensively discussed and analyzed in Zhang and Jacobsen (2013) and they find it is well suited for studying monthly anomalies as we do here, even though dividends are excluded, the series combines value and equally weighted indices and is initially only based on three different stocks. We refer the reader concerned about the possible impact of these effects to their extensive robustnes checks. In short they find that when the index consisted of three stocks, these stocks essentially were the complete market in the early days of the stock exchange and this index highly correlates with another (annual) index based on eight stocks created by Mirowski (1981) during a large part of that sample. They also find no evidence of clustering of dividends in specific months. And they show if they replace indices as much as possible with value weighted indices, results tend to be similar.

US stock return index

We have chosen the UK price return index due to the length of the available data. However, our methodology is applicable to a variety of data series. US stock returns are used most frequently in asset pricing research. Therefore, it is natural for us to apply test our methodology on the US sample.

US price return index is available from Global Financial Data from 1792, giving us 221 years of data. Although not as long of a series as UK return series, US data could give us additional insights.

3. Methodology

Detecting and measuring statistical signals

In a nutshell we test how forecasting strength varies when we vary estimation intervals and confidence levels, using different assumption for the underlying data generating process. We now describe this approach in more detail.

Note that we are not considering investment decisions as such. We are considering whether analyzing the statistical properties of a time series (which happens to be an over three hundred year long time series of monthly observations of a stock market index) has forecasting ability. This means that someone must be able to detect some sort of signal from the series. This signal would need to give valuable information regarding the future.

A simple approach would be to assume that an investor assumes there is no forecastability other than saying that the stock market may increase on average. In return predictability research this is often the benchmark against which models of stock market predictability are tested.

In that case the investor expects the level of the index at the end of next month to be equal

$$E[P_t] = P_{t-1}(1 + \mu^B) \qquad (1)$$

and thus just increases with a constant factor μ^B on average. We use the superscript B to indicate that this will be our benchmark against we will measure alternative forecasts.

Or, in returns, $E[r_t] = \mu^B$. Essentially this investor assumes stock market returns follow a simple random walk (with a drift).

Alternatively, one might consider that return in the month August might be significantly higher or lower than this mean return. In that case we would get the model:

$$E[r_t] = \mu^A + \alpha AUG_t \quad (2)$$

Where AUG_t is a dummy variable that takes the value 1 if a month is August and 0 otherwise and α denotes the deviation from the monthly mean return and a positive α would indicate that returns were indeed higher. The superscript *A* to denotes the alternative mean return To determine reasonable parameter estimates for these expectations our statistician will have to assign values to the parameters or estimate them based on a historically relevant period.

If the investor feels this is an appropriate model to describe the past we can simply transform this equation into a regression by adding realizations of the returns on both sides, subtracting expectations on both sides and we get:

$$r_t = \mu_T^A + \alpha_T AUG_t + \varepsilon_t \text{ with } \varepsilon_t = r_t - E[r_t] \quad (3)$$

We use *T* to indicate the estimation window used to estimate the parameters μ and α. This means that to extract signals based on this model, the investor will need to pick:

1) some measurement interval *T*;
2) some significance level (*p*-value);
3) some model with assumptions for the underlying stochastic process.

Determining forecasting success

We measure forecasting accuracy based on whether returns in the month August where indeed higher than the benchmark average μ_T^A implied by the benchmark model.

Thus an investor considers wanting to predict the market in August 2013, looks back *T (e.g, 35)* years ending July 2013. Based on a confidence level *c (e.g., 5%)* determines whether α_T is significant (positive or negative). If average returns in August (measured over *T (35)* years) are significantly positive (negative) the investor predicts higher (lower) than average returns. If August 2013 is indeed higher (lower) than the mean returns over August 1978 – July 2013, the signal was right and we call this a correct forecast. Then we move on to the next month. We verify whether there is a September effect measured over *T* years before September 2013 and so on. One investor may think that the world changes fast and that with those the underlying return generating process changes fast as well. Other investors might argue that

the more the world changes the more it stays the same and argue for a longer measurement interval to get parameter estimates. We vary *T* and *c* and our estimation methods to test whether there are optimal levels for these variables and study how these are affect results.

We consider all months but exclude January as a January effect has been well documented in the literature and is observed in this series since 1835 (see Zhang and Jacobsen, 2013). This January hindsight bias may positively affect our percentage of correct predictions. Including January in the estimations does not dramatically alter the findings. This is discussed in detail in Section 5.

4. Results

4.1 Main result

As a baseline specification, we consider an investor who detects signals using OLS regression with White standard errors at 5% significance level. We begin with this specification as it is one of the most commonly used in empirical finance. Alternative significance levels and model specifications are considered in sections that follow.

Specifically, the investor runs the following OLS White regression:

$$r_t = \alpha + \beta Feb + \varepsilon_t \qquad (4)$$

for the period of T years prior to any given February. The regression is then run for March-December dummies. The results are presented in Figure 1.

[INSERT FIGURE 1 HERE]

We observe a number of things. First, forecasting accuracy is not incredibly high. The average success rate for all detection window lengths is 55.61%. Second, the choice of an "optimal" detection window length is far from clear. At extremely low window lengths, success rates behave rather erratically, with a somewhat upward trend. Best success rates appear to be achieved with detection window lengths of about 23 – 37 years. As a matter of fact, the highest success rate of 62.18% is achieved with 35 years. After that, there is a fairly linear downward movement in success rates, and they actually dip below 50% for detection periods of about 170 years.

Things are even less clear at longer window lengths. We observe a somewhat upward trend for window lengths between about 175 and 265 years, after which the pattern becomes erratic

again. It is important to note that when we use extremely long detection windows (the maximum length is 300 years), we do not have many observations left for out of sample estimation of signals' predictive ability (only 17 years for the 300 year detection window). Therefore, predictive abilities of extremely long detection windows must be interpreted with caution.

4.2 Impact of significance levels

The results above assume that the investor "extracts" the signal with a 5% significance level – the level most commonly used in empirical finance. But what if higher/lower significance levels result in better forecasting ability? We extract signals at 1, 5, 10, 20, 30, 50, and 99% significance levels. Figure 2 presents the results.

[INSERT FIGURE 2 HERE]

The results are surprising to say the least. It is intuitive to expect signals detected with higher significance to result in better forecasting ability. After all, this is the whole premise of return predictability. However, only using reasonably short detection windows (up to 120 years) signals detected with 1% significance tend to show the highest forecasting accuracy. Interestingly, these signals display the *worst* forecasting accuracy at the detection window lengths between approximately 205 and 240 years.

Figure 3 presents "zoomed in" results using the detection windows between 3 and 90 years, which, in many practical applications, are more reasonable detection windows.

[INSERT FIGURE 3 HERE]

Signals detected with 1% significance level tend to have better forecasting accuracy than signals detected with lower levels, albeit with higher volatility. The difference between signals detected with 5% and 10% levels is, however, negligible.

Rather than comparing the signal's forecasting ability to a 50% benchmark, it could be more instructive to compare forecasting success of various significance levels to a "baseline" case of a 99% significance. Up to this point, we have assumed that the benchmark forecasting accuracy is 50% - a pure random walk. We now assume that a signal – positive or negative – is detected every month. As evident from Figures 2 and 3, forecasting accuracy of such signals is almost always greater than 50% (on average, the forecasting accuracy is 52.44%). We present the results in Figure 4.

[INSERT FIGURE 4 HERE]

A number of interesting observations can be made. On average, signals detected with 1% accuracy outperform the benchmark (99%) by 4.17% in terms of forecasting accuracy. These signals also show the highest outperformance of 14.02% (when a 30-year detection window is used). However, 1% signals also have the worst underperformance of all – -8.02%. These signals perform poorly at detection windows of around 220 years. As for the 5% signals, their average outperformance is 3.18%, with peak outperformance of 8.70% and worst underperformance of -5.75%. For 10% signals, the numbers are 2.04%, 7.05% and -3.90% respectively. While average outperformance is indeed higher for signals detected with higher significance, their underperformance is worse as well. Also, all significance levels considered produce the best forecasting ability at some length of the detection window.

4.3 Choice of estimation methodology

To this point, the investor was extracting signals using OLS regression with White standard errors. In this section, we evaluate the impact of estimation methodologies on signals' forecasting ability. We consider the following methodologies:

- T-test of a difference between a given month's return and zero;
- Paired-sample t-test of difference between a given month's return and average return for the calendar year;
- Simple OLS with no White standard errors;
- Robust regression;
- Estimation using GARCH(1,1) residuals;
- Probit estimation. The dichotomous dummy variable was equal to 1 if a given month's return was above/below the calendar year's mean return.

The results are presented in Figure 5.

[INSERT FIGURE 5 HERE]

Several observations can be made. First, there is no clear "winner" in terms of forecasting success – the results are highly dependent on the length of the detection window. Second, the simplest "zero t-test" has the highest average forecasting success rate of 57.46%. The average rates for other methodologies are as follows: 55.43% for a paired-sample t-test, 55.24% for OLS, 54.53% for GARCH, 54.43% for Probit and 53.87% for robust regression.

It is important to note that although every methodology relies on reasonable number of observations for reliability, the issue is particularly relevant for GARCH and Probit. Therefore, we used a minimum of 30 years for detection window lengths when detecting signals using these methodologies.

5. Robustness checks

5.1US stock returns

As US data is most commonly used in empirical research, we apply our methodology to the US stock market as well. Here, due to the shorter series (221 years), our maximum detection window is 200 years. The results are presented in Figure 6. Note that the two graphs compare signals detected with a 5% significance level.

[INSERT FIGURE 6 HERE]

The forecasting success of signals is substantially lower than in the UK market. Average forecasting success is 51.05%, as opposed to 55.61% for the UK. However, we observe certain similarities between the US and the UK markets. First, forecasting accuracy is at its highest for relatively short detection windows. Second, there appears to be a downward trend in forecasting accuracy as the detection window length increases, bottoming out at less than 40% with the detection window of 127 years. After that, forecasting accuracy behaves rather erratically, perhaps due to a relatively small number of signals.

5.2 Starting in the 20th century

Stock market returns were fairly flat before the beginning of the 20th century (some may argue that it wasn't till the end of WWII that stock prices have started rapid growth). Therefore, we apply the methodology to US and UK stock return series starting in 1900 giving us 113 yearly observations. We thus limit the detection window to a maximum of 100 years. The results for signals detected with a 5% significance are presented in Figure 7.

[INSERT FIGURE 7 HERE]

We observe that, with few exceptions, forecasting ability in the UK sample is greater than in the US sample. Average forecasting success for UK is 60.67%, and for US is 53.43%. In both cases the forecasting ability is substantially higher than in the "overall" sample, which may imply that from 1900 the stock return series are actually further from random walk.

We also observe a fairly consistent upward trend in forecasting success for both series, implying that signals detected with longer windows have better predictive abilities. This is somewhat surprising. When deciding which detection window length to use, one faces a trade-off between higher accuracy of the signal (i.e., signal based on more observation) and the fact that the underlying data-generating process may change. It is reasonable to suspect that the flow of information starting from 1900, as well as the number of potentially regime-shifting events has increased in frequency. Thus, we may expect signals based on *shorter* detection windows to have better forecasting success. However, the highest forecasting success rate for the UK sample is achieved for the detection window of 90 years (with a maximum window length of 100). For the UK, the most successful detection window is 96 years.

5.3 Including the January effect

As January effect has been documented in the literature and could argue that our methodology is just another way to detect the January effect, albeit with more noise. Thus, we have excluded the month of January from the analysis. But what happens if we don't? The results are presented in Figure 8.

[INSERT FIGURE 8 HERE]

Forecasting abilities of signals improve a bit for virtually all detection window lengths. However, the overall pattern in forecasting success rates as a function of detection window lengths remains virtually unchanged. This leads us to believe that our findings are not driven primarily by the January effect.

6. Conclusion

What is the optimal number of observations to use for predictive purposes? Are predictions based on signals detected with higher significance and complex techniques produce better results. We provide somewhat surprising answers to both questions.

First, the optimal detection window length for monthly signals appears to be rather low. This suggests that the changing nature of the return generating process outweighs accuracy gains achieved with longer detection windows. Another important implication is that the number of signals (and thus potential trading opportunities) is higher with shorter detection windows. Therefore, an investor may not need to wait vary long to establish the signal.

Second, signals detected with higher levels of statistical significance do not necessarily imply higher forecasting successes. Using complex methodologies for signal detection does not improve forecasting accuracies either.

Our research question opens ample opportunities for future research as it is applicable to a variety of issues in empirical finance and beyond.

References

Anserson, M.J. & Thompson, A.A. (2004). Multivariate control charts for ecological and environmental monitoring. *Ecological Applications* 14(6), 1921-1935.

Cenesizoglu, T. & Timmernan, A. (2011). Do return prediction models add economic value? *Working Paper*.

Driesong, G., Jacobsen, B. & Maat, B. (2008). Striking oil: Another puzzle? *Journal of Financial Economics* 89, 307-327.

Pesaran, M.H. & Timmerman, A. (1995). Predictability of stock returns: Robustness and economic significance. *Journal of Finance* 50(4), 1201-1228.

Rapach, D.E. & Wohar, M.E. (2006). In-sample vs. out-of-sample tests of stock return predictability in the context of data mining. *Journal of Empirical Finance* 13, 231-247.

Timmerman, A. (2007). Elusive return predictability. *Working Paper.*

Wachtel, S. B. (1942). Certain observations on seasonal movements in stock prices. *Journal of Business, 15*, 184-193.

Zhang, C.Y. & Jacobsen, B. (2012). Are monthly seasonal real? A three century perspective. *Review of Finance,* forthcoming.

Figure 1. Forecasting ability of signals.

The figure presents forecasting ability of signals at various detection window lengths. Signals are detected by OLS regressions with White standard errors at a 5% significance level.

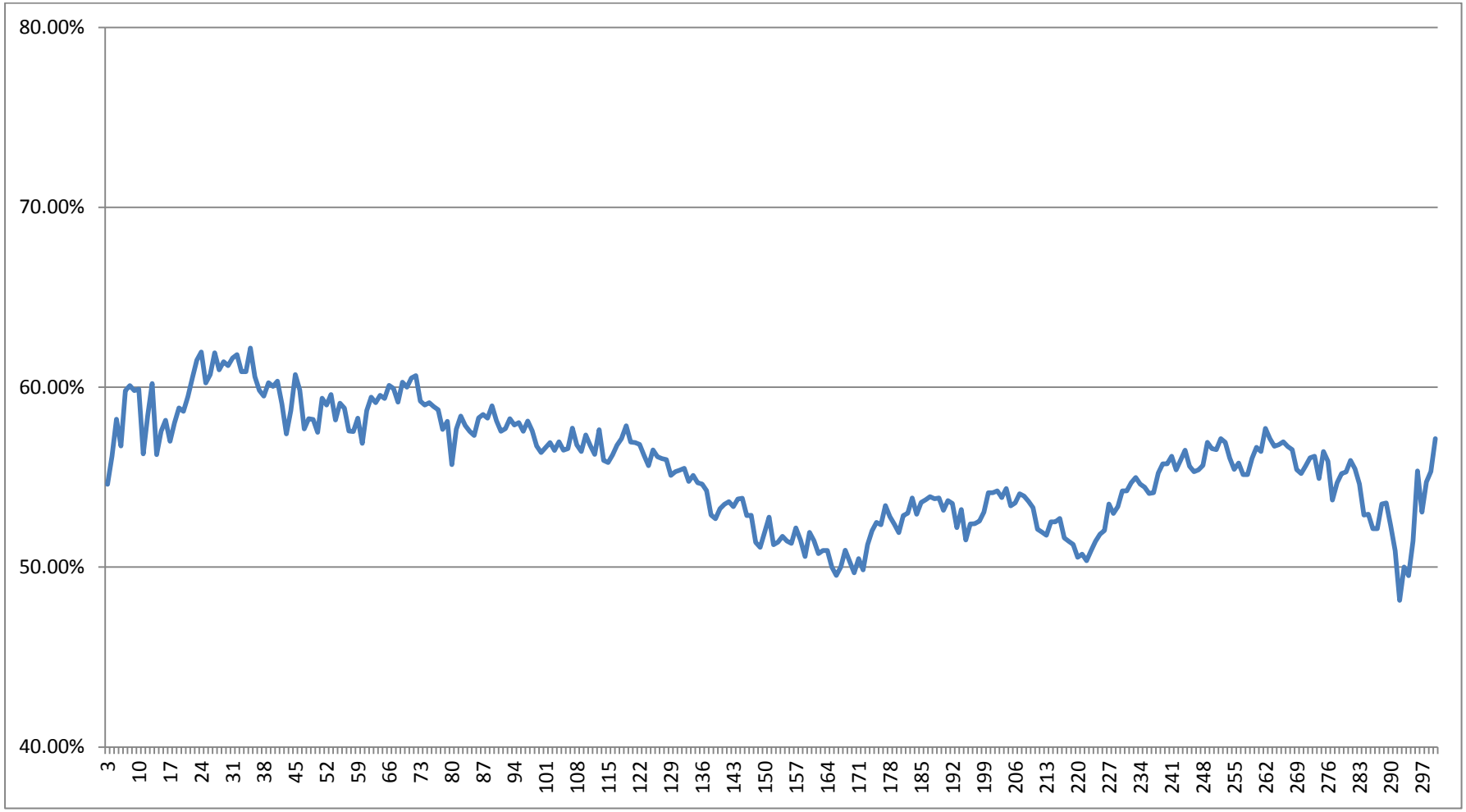

Figure 2. Forecasting ability of signals at various significance levels.

The figure presents forecasting ability of signals at various detection window lengths. Signals are detected by OLS regressions with White standard errors. Various levels of statistical significance are used for signal detection.

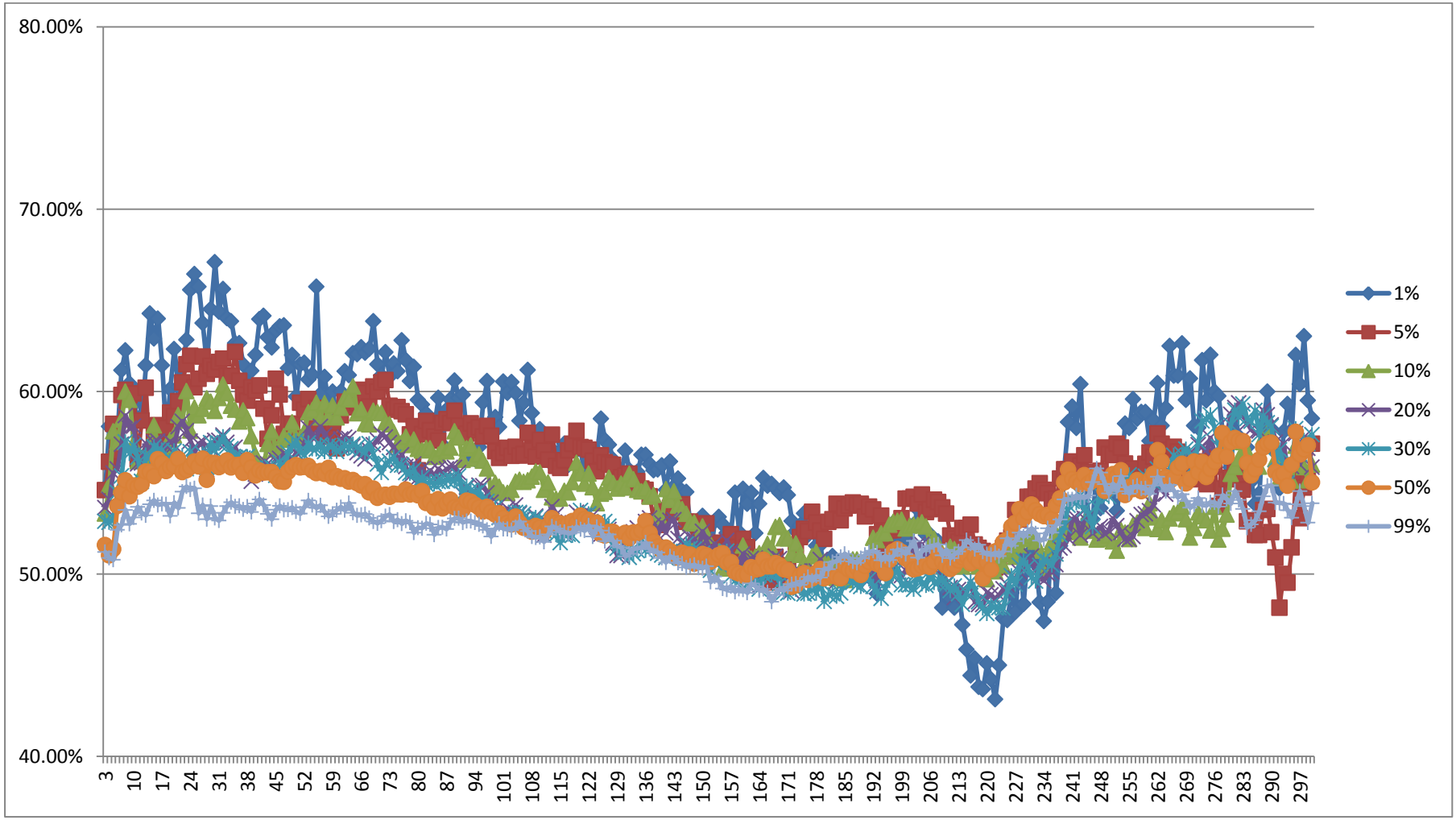

Figure 3. Forecasting ability of signals using shorter detection windows.

The figure presents forecasting ability of signals (detected using various significance levels) for the detection windows of up to 90 years.

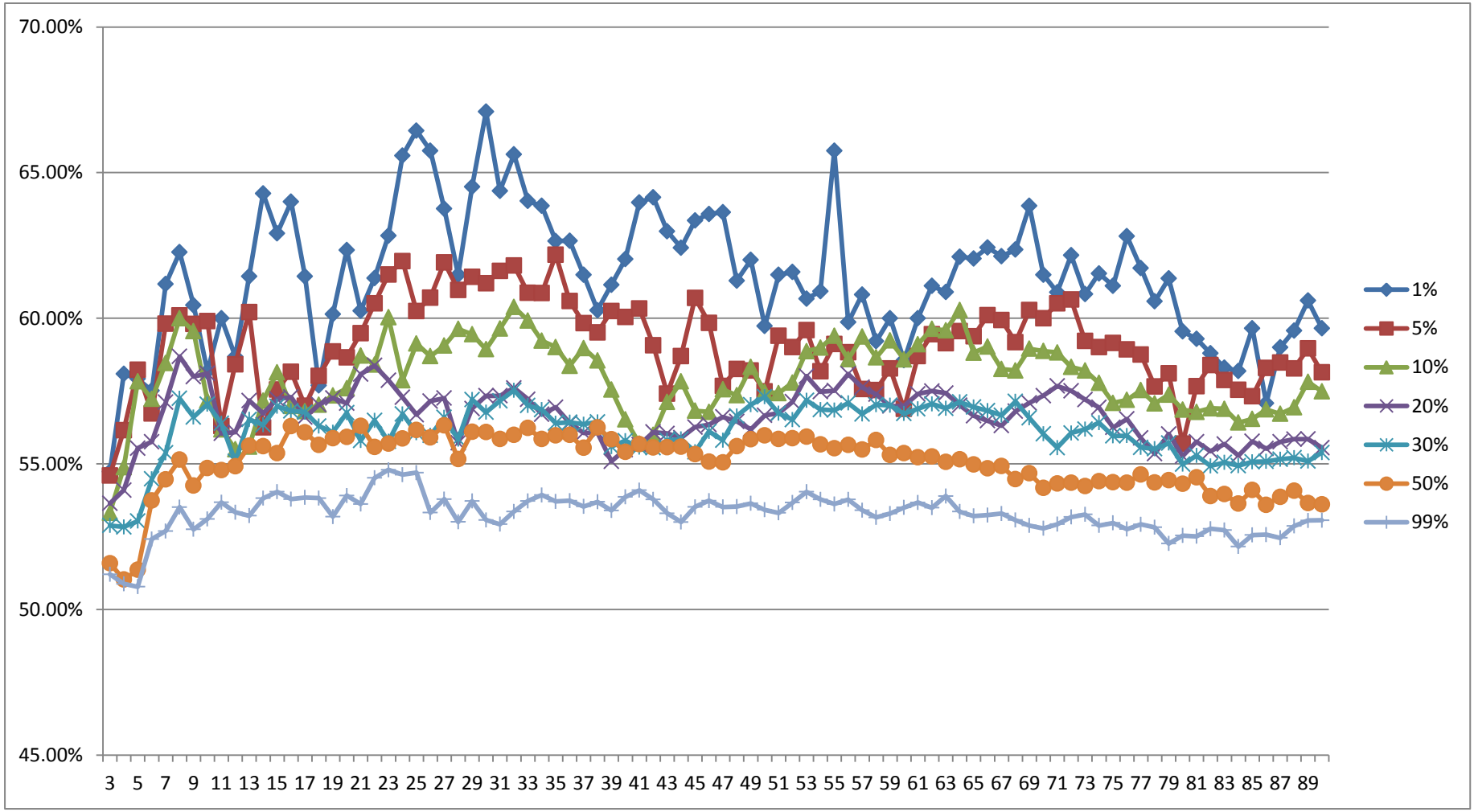

Figure 4. Forecasting ability relative to a 99% benchmark.

The figure represents over/under performance of signals detected with various significance levels relative to signals detected with a 99% significance level.

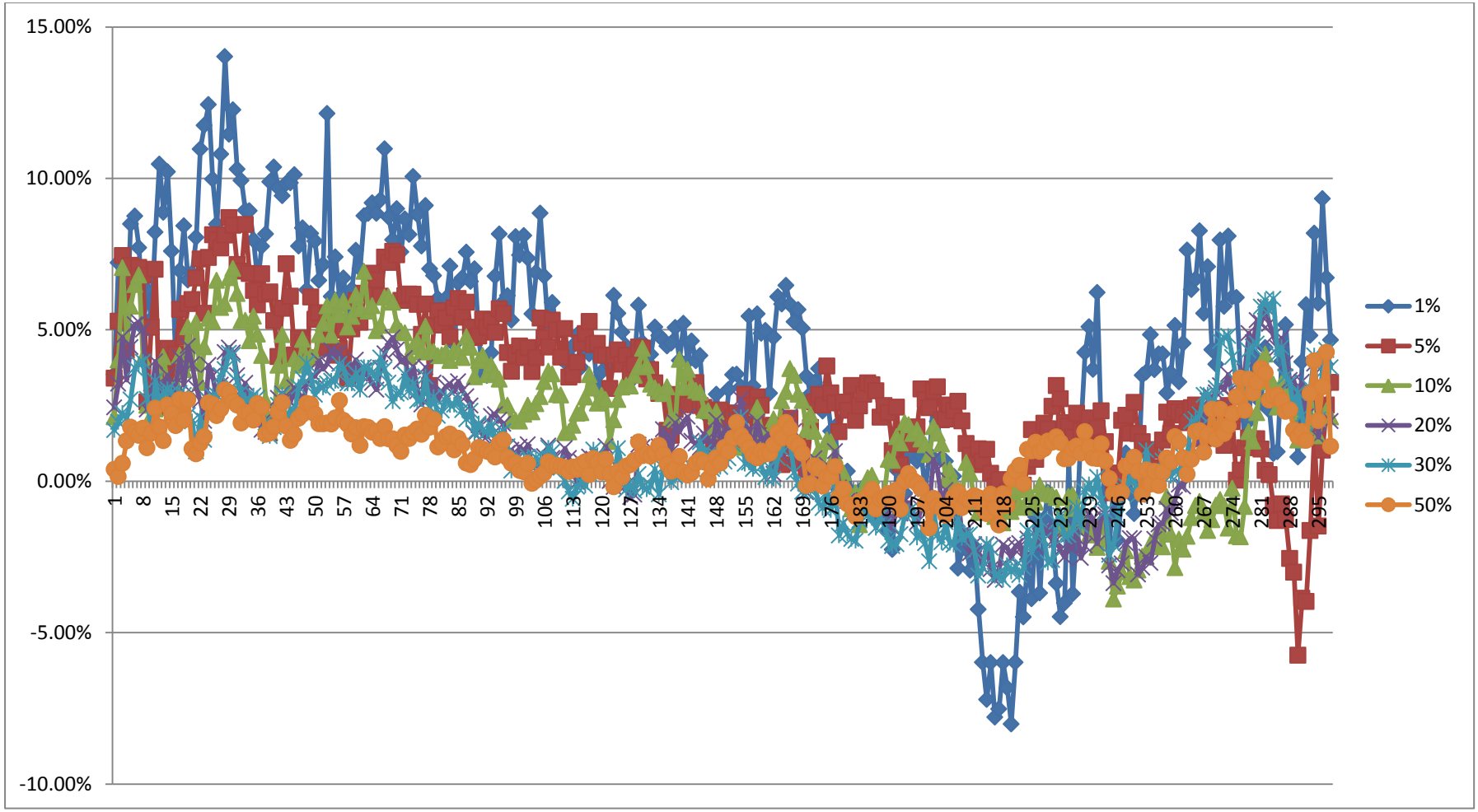

Figure 5. Methodology Comparison.

The figure presents forecasting success of signals detected using various methodologies. 5% significance levels were used for all methodologies.

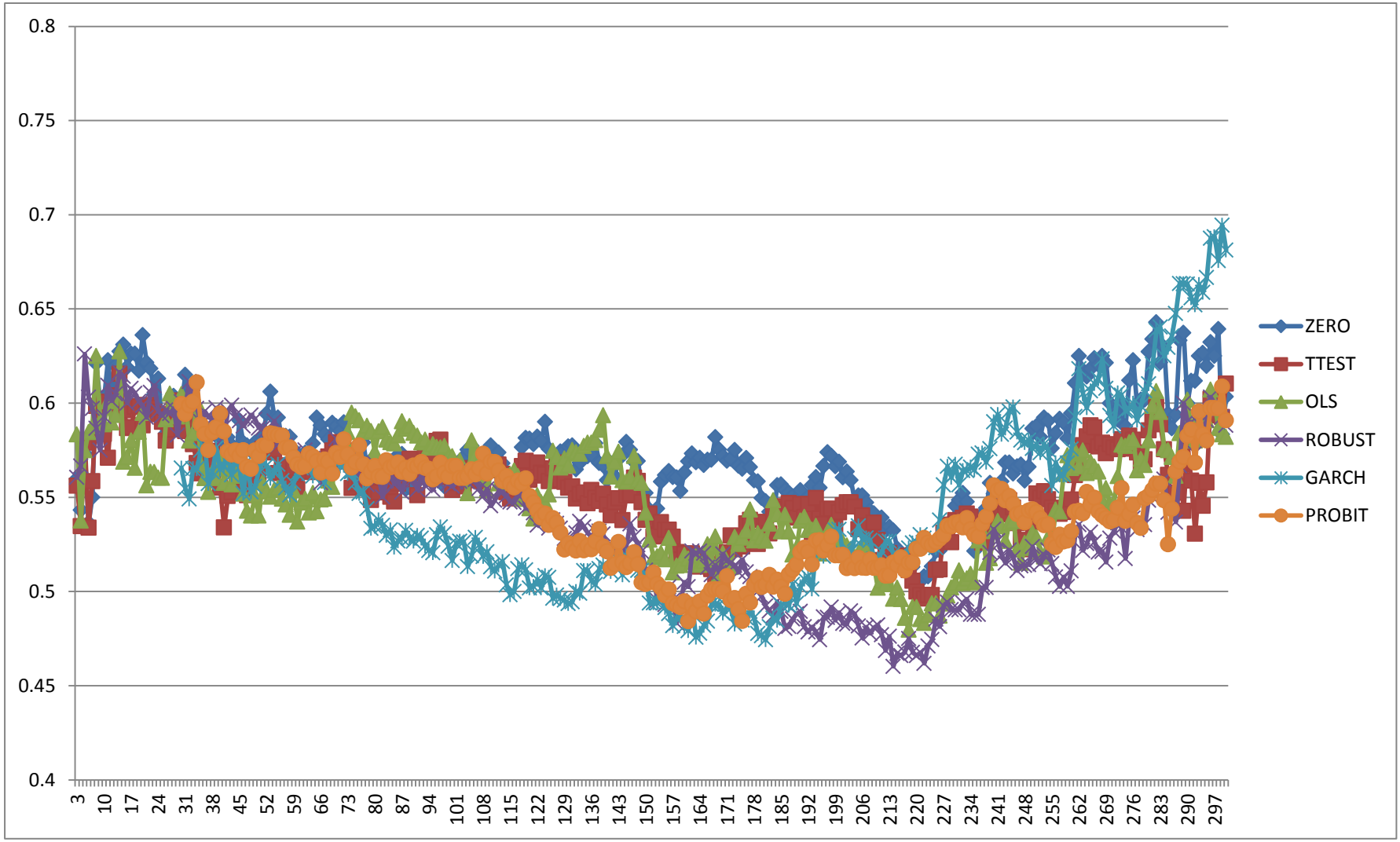

Figure 6. US data series.

The figure presents forecasting ability of signals at various detection window lengths using US data. Signals are detected by OLS regressions with White standard errors at a 5% significance.

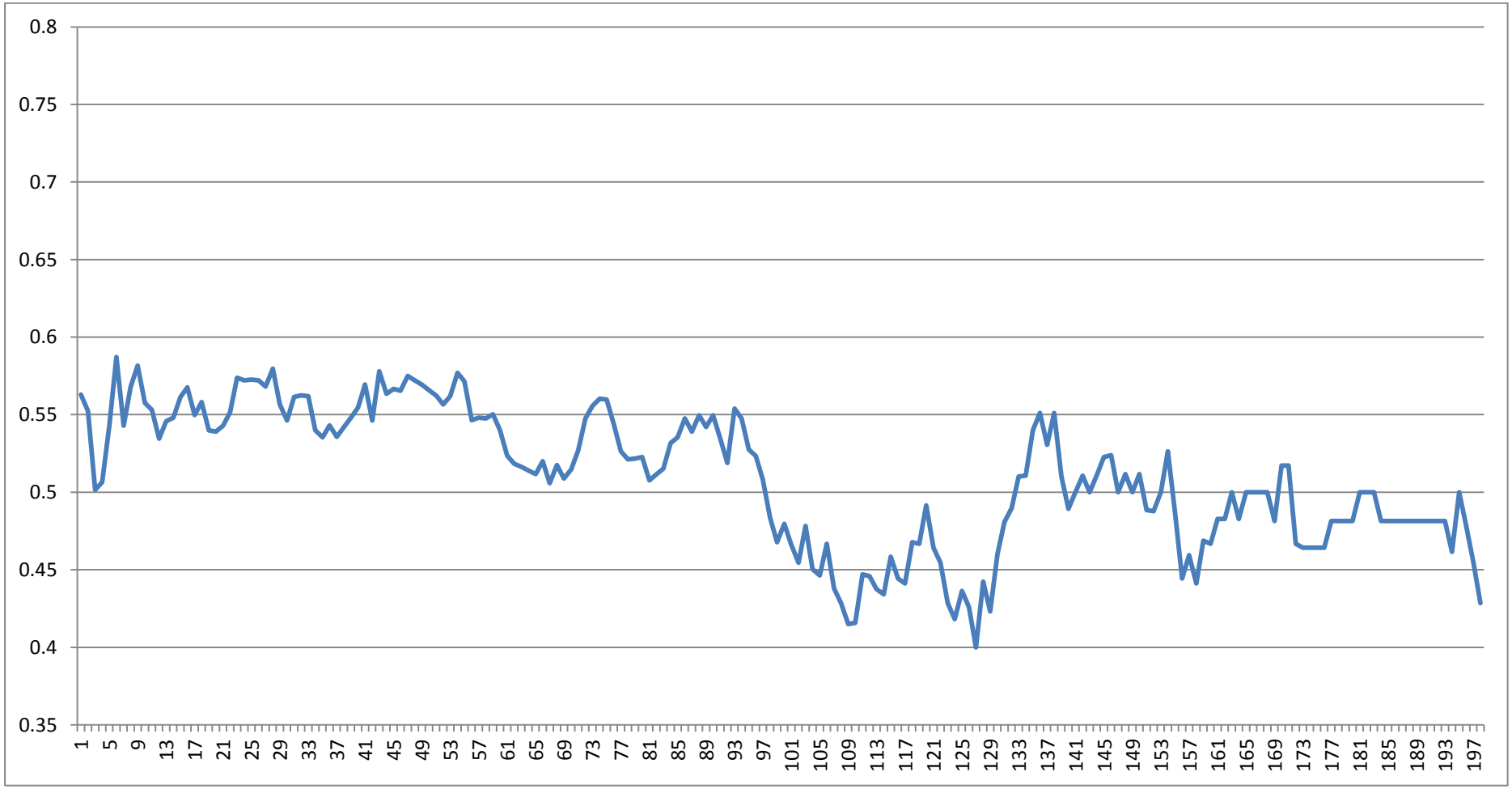

Figure 7 – US and UK data from 1900.

The figure presents forecasting success of signals for both US and UK series starting in 1900. The signals are detected using OLS with White standard errors at a 5% significance level.

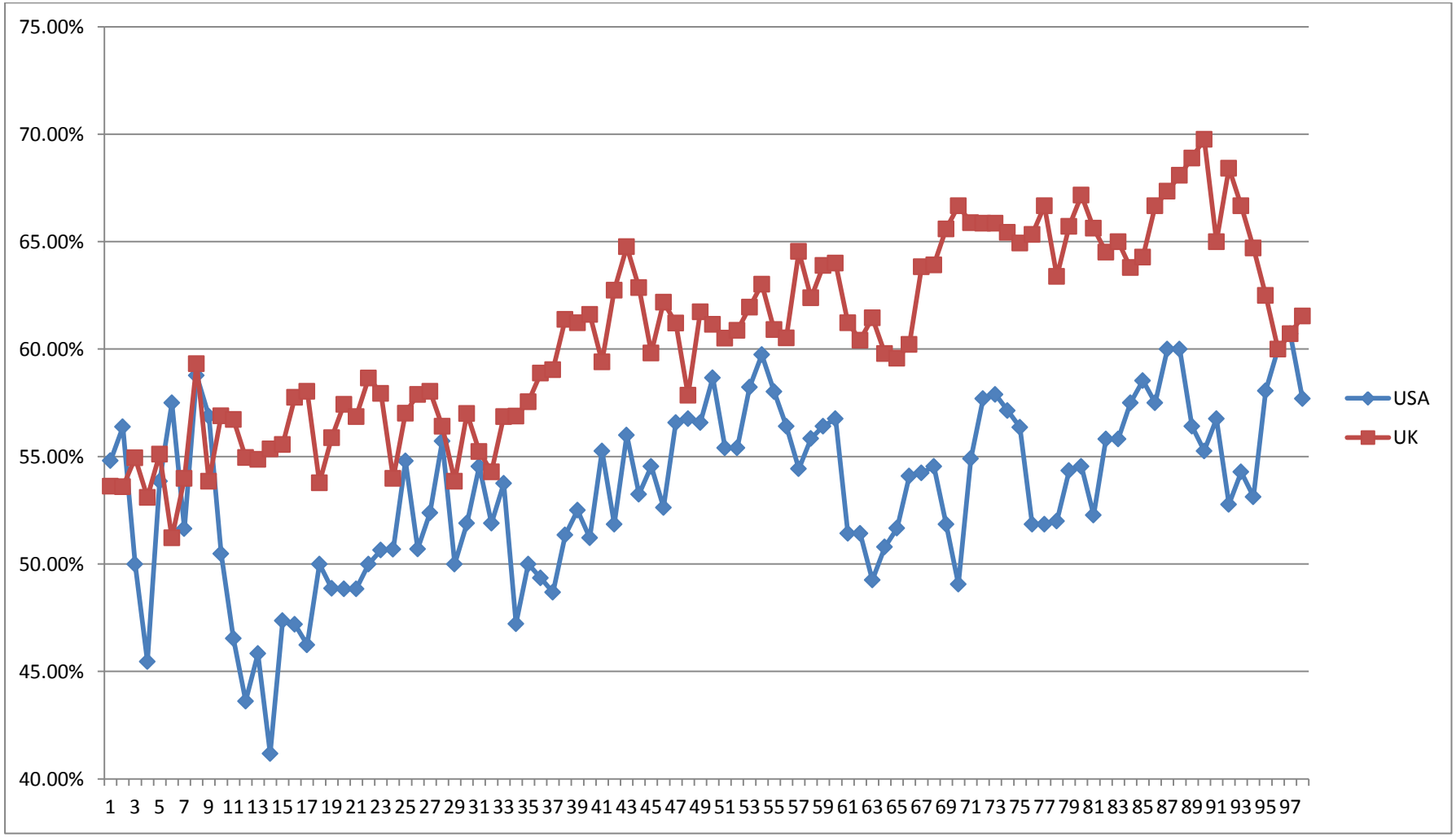

Figure 8. Forecasting ability of signals (including January)

The figure presents forecasting ability of signals at various detection window lengths. Signals are detected by OLS regressions with White standard errors at a 5% significance level. The month of January is included in the sample.