How Good Is a "Good" Forecast?: Forecast Errors and Their Avoidability

STEVE MORLIDGE

PREVIEW *With this article,* Foresight *continues its examination of* forecastability *– the potential accuracy of our forecasting efforts – which is one of the most perplexing yet essential issues for the business forecasting profession. We first tackled the subject with a special feature section in our Spring 2009 issue. My introduction there indicated that assessing the forecastability of a historical time series can give us a basis for judging how successful our modeling has been (benchmarking), and how much improvement we can still hope to attain.*

Foresight*'s Summer 2012 issue advanced the discussion with a feature article showing how to use a product's DNA – product and market attributes such as the length, variability, and market concentration of sales – to develop benchmarks for forecast accuracy. The essential idea here is to better understand the specifics of those items we are trying to forecast and to set expectations accordingly.*

Certain key concepts emerged from the articles in that section that helped clarify the meaning of forecastability and the challenges underlying its analysis:

- *The lower and upper bounds of forecast accuracy – the worst and best accuracy to be expected*
- *The relationship between the volatility of our sales histories over time and their forecastability*
- *The limitations of the coefficient of variation in measuring forecastability and a potentially better alternative in a metric of entropy*

Now Steve Morlidge offers a tantalizing new perspective on forecastability. His approach seeks to determine what portion of forecast error for any item is avoidable, in principle and in practice. The simplicity of the metric he creates should be very appealing to business forecasters, seeing that it offers a convenient way to compare accuracy results across products.

BEGINNINGS

It all started in 2004.

I was working in a large multinational company, responsible for developing and promoting a performance-management initiative in the finance function. The books on managing change that I had been reading made it clear that bringing about change depends on having a "mission critical" problem – a burning platform – and identifying what you were doing as the solution.

It was clear to me that our financial forecasting was a broken process. I needed to spur people into action, and I had spent over a year working up and promoting a solution to the problem. And then – to my good fortune, if not that of the shareholders – my company was forced to deliver the first profit warning in its proud history.

In a matter of weeks, I found myself at the heart of efforts to fight the fires that broke out across the business as a result of this public admission

Special Feature

Key Points

- While it may never be possible to determine the best accuracy one can hope to achieve in forecasting any particular item, we can demonstrate what level of forecast error is *unavoidable* – a significant step toward being able to make objective statements about forecast quality.

- What proportion of the error is avoidable? In principle, *bias* – the tendency of a forecast to systematically miss the actual demand (consistently either high or low) – is avoidable, but some portion of the *error magnitude* is unavoidable because there will always be an element of randomness in our data sets.

- A simple but effective way to measure that unavoidable portion is on the basis of the forecast errors from a *naïve model*, which issues forecasts of "no change" from the present to the future.

- While this is not a new idea, we show that, under common circumstances, ratios of the forecast errors from your model to those of a naïve model have *natural lower bounds*, which provide benchmarks for seeing if you have eliminated all but unavoidable error.

of failure. My first step was to draft a forecast policy, the reason for which was simple: like most other companies, my employer had never formally defined what a good forecast should look like. Without a definition of success, it was little wonder that our forecast processes had failed so catastrophically. Fortunately, I had prepared myself well for this task.

Defining Success in Forecasting

In my research of the previous year, I had discovered that the science of forecasting in finance was primitive in the extreme. No one in the field seemed to have a clear idea about what constituted a good forecast. As fortune would have it, I had attached myself to a group that had been working for a number of years to improve planning and forecasting practice in the supply chain, and I learned a great deal – not all of it for the first time – that I was able to use in my developing ideas about how finance should go about things. The definition of success that our group used was this:

"A good forecast exhibits no bias and minimal variation."

This definition correctly recognises that systematic error (bias) and unsystematic error (variability or volatility) have different characteristics and consequences for the business. With a rapidity that was all but unprecedented, our definition of success (with a few tweaks to accommodate the peculiarities of financial forecasting) was adopted as a corporate policy.

Afterward, the company finance team with which I'd developed the new forecast policy invited me in for celebratory tea and biscuits. As we chatted, one team member asked me casually enough, "This is great, Steve, but how do we know if we have actually got a good forecast?"

Try as I might, I had no answer. The best I could do was, "Good question. Leave it with me." Like many simple questions, it was not as easy to answer as it perhaps first appeared.

Creating a Metric

Over the next few months, I was forced to come to terms with the subtlety of the problem and the depth of my ignorance on the subject. I formed a clear view of what kind of measurement system we needed to operationalize the policy that I had helped draft:

- It should be able to distinguish forecast error bias from forecast error magnitude (i.e., unsystematic variation).

- It should be actionable; being "accurate enough and quick" was better than "perfect and slow," since we needed to correct problems before they had a chance to overwhelm us.

- It had to recognize the difference between signal and noise; that is, it should alert us to real problems and deter us from intervening when there was no evidence of an issue problem.

- It should be simple to calculate and easy to communicate to non-experts.

- It would quantify what constitutes an *acceptable level of forecast accuracy*.

I slowly came to understand that this final criterion presented the most formidable obstacle because it had three distinct facets:

1 **How forecastable is the data set?** Clearly, we cannot expect the same degree of error for a low-level forecast in a volatile market as for a high-level forecast in a stable market.

2 **What proportion of the error is avoidable?** *Bias*, the tendency of a forecast to systematically miss the actual demand (consistently either high or low), is avoidable in principle – but some portion of the forecast error is unavoidable because there is always going to be an element of randomness in our data. It is true that biases can arise after a major structural change, but a good forecasting algorithm should be able to detect systematic error and correct for it before it builds up.

3 **What is the business impact of the forecast error?** For example, we might be happy to tolerate a high level of errors where the impact (in terms of cost of inventory, for example) is relatively low.

Unsurprisingly, these same questions have exercised the best minds in our field, as a review of past issues of *Foresight* makes abundantly apparent.

WHAT THE EXPERTS SAY

There is arguably no topic in forecasting more passionately debated than that of forecastability.

The most widely promoted approach is based on the intuitive insight that, generally, the more volatile the variable, the more difficult it is to forecast. There is a large body of empirical support for this concept. The Coefficient of Variation (CoV) – the ratio of the variation from the average in the data to the average value – is a standard measure of variability. Thus researchers have sought to correlate forecast accuracy with the CoV (Gilliland, 2010, Schubert, 2012).

One shortcoming with the CoV is that it does not always correlate well with forecast accuracy (Schubert, 2012); and even if it did measure actual forecast accuracy, it would not necessarily reflect forecastability (potential forecast accuracy).

Popular alternative approaches are based on comparisons of forecast accuracy with a benchmark such as the accuracy of a naïve forecast, where the actual for a period is used as the forecast for the subsequent period. Metrics employed in this approach are ratios of forecast errors from a designated model to the naïve forecast errors, and include Theil's U statistic (1966), the Relative Absolute Error or RAE (Armstrong and Collopy, 1992), the Mean Absolute Scaled Error or MASE (Hyndman, 2006), as well as the concept of Forecast Value Added (Gilliland, 2013).

An advantage of using the naïve forecast as a benchmark is that it implicitly incorporates the notion of volatility, since the naïve forecast has the same level of variation as the variable itself. Errors associated with the naïve forecast are also probably a better predictor of forecastability for time-series purposes than the Coefficient of Variation because they measure period-to-period variation in the data. For example, a series where successive observations are highly positively correlated (so the series is forecastable) may drift away from the series' mean for several periods, thereby contributing to a high CoV. In contrast, the naïve forecast errors will be relatively small because the successive observations are similar.

A number of authors have expressed discomfort with using any forecast accuracy metric as a proxy for forecastability (Boylan, 2009). Peter Catt demonstrated (2009) how completely deterministic processes – and thus totally forecastable if you know the data generating process – can create very volatile data series. Attempts to find ways to measure forecastability directly have foundered on the self-referential nature of the problem: we can only assess the performance of a forecasting methodology by comparison with an unspecifiable set of all possible methodologies.

These authors have proposed alternative ways of assessing forecastability, such as through a profile of a "product DNA" (Schubert, 2012). It comes as no surprise that these methods are relatively complex and consequently more difficult to implement and interpret. A more straightforward approach emerges from the concept of *avoidability*.

AVOIDABILITY

Avoidability is closely related to forecastability. John Boylan (2009) defines forecastability as "the range of forecast errors that are achievable on average, in the long run." He argued that the upper bound of forecast error should be the naïve forecast error. This is an uncontroversial position since the naïve is the crudest forecast process imaginable – albeit one that professional forecasters often fail to beat in practice (Pearson, 2010). The lower bound or lowest achievable forecast error, Boylan indicates, could be impossible to determine because there are "endless forecasting methods that may be used. It is possible that a series is difficult to forecast and will yield high forecast errors unless a particular method is identified."

Avoidability sets a theoretical lower bound to the forecast error that is independent of the forecaster and the available tool set, and it can be quantified using a common error metric such as Mean Squared Error (MSE) or Mean Absolute Error (MAE). The theoretical lower bound may be achievable only with tools beyond the reach of the forecaster. What is achievable using existing technology defines *forecastability*.

What I was attempting to do all those years ago – without realising it – was to build a forecasting control system. I have learned since I embarked on this quest that, without good feedback, no process can be relied upon to consistently deliver a desired output. This fact surrounds us in nature, and it is at the heart of all of mankind's technological advances. Our bodies regulate the levels of many thousands of chemicals in a way that is very similar to how modern engine-management systems optimise the performance of our motor vehicles. In the same way, no forecast methodology, no matter how sophisticated, can consistently deliver a good performance unless we can find a way to measure and compare its performance to the desired result. Doing so enables us to make the timely corrections necessary to eliminate unnecessary and unwanted error (see Hoover, 2006).

It appears, then, that being able to determine what level of performance is achievable is not the icing on the forecasting cake after all; it is the difference between interesting mathematical theory and useful technology. Finding a way to break though the complexity surrounding these issues is imperative. Fortunately, recent work has suggested an approach.

The Way Forward: A Conjecture

In attempting to understand what constitutes an acceptable level of forecast performance, we start with these standard assertions:

1. First, there are no conceivable circumstances where forecasting performance should be consistently worse than that of the naïve forecast.
2. Second, the performance of any system that we might want to forecast will always contain noise.

With regard to number 2, we know that all extrapolation-based forecasting (i.e., time-series forecasting) rests on the assumption that there is a pattern (or signal) in the past data that will influence future outcomes, and that this signal is obscured by randomness. In addition, we should always expect that the signal will change at least a little bit as we move into the future – just how and how much are unknowable at present. So the job of a forecasting algorithm is to detect and mathematically describe the past pattern – having excluded the noise – and then apply it to extrapolate into the future.

A "perfect" forecasting algorithm would describe the past signal, leaving only errors that represent pure noise, and hence unavoidable. Since the errors from a naïve forecast are one way of measuring the observed amount of noise in data, my conjecture is that there is a mathematical relationship between these naïve forecast errors and the *lowest possible errors* from a forecast.

The Unavoidability Ratio

Prompted by this conjecture, Paul Goodwin (2013) provides a mathematical derivation of what this relationship might be. We summarize the results here:

When the pattern in the data is purely random, the ratio of the variance (mean squared error, MSE) from a perfect algorithm to the MSE of a naïve forecast will be 0.5; that is, the perfect algorithm will cut observed noise (using the MSE measure) in half. Using the more practical measure of the ratio of the mean *absolute* error (MAE),

THE ASSUMPTIONS:

- We have the perfect forecasting algorithm.
- The remaining errors are pure noise in the statistical sense that they are "stationary and independently and identically distributed with a mean of zero."
- The change in the signal from period to period is unaffected by the previous period's noise.

THE UNAVOIDABILITY RATIO

Under these assumptions, the ratio of the variance of pure error (that is, error from a perfect forecasting algorithm) to that of the errors from a naïve forecast model will be:

$$\frac{\text{Variance of pure error}}{2\,(\text{Variance of pure error}) + \text{Variance of period-to-period changes in signal} + (\text{Mean change in signal})^2}$$

If there are no systematic changes in the signal (e.g., no trend or cyclical pattern), the second and third terms in the denominator become zero, leaving us with

$$\frac{\text{Variance of noise}}{2\,(\text{Variance of noise})} = 0.5$$

for the best possible performance, and thus the definition of what constitutes unavoidable error.

a "perfect" algorithm would never achieve a ratio lower than 0.7 (√0.5).

This surprisingly simple result emerges from a particular set of assumptions about the data, which we enumerate in the accompanying boxed inset. The key assumption is that there is no trend, cyclical pattern to the historical data, or impact from causal variables.

Some might argue that this approach has limited value since it is not safe to assume that there will be no systematic changes in the signal; the existence of anything other than a flat trend, particularly if nonlinear, could lead to much lower theoretical minimum. However, there are many real-life situations where our assumptions can apply. For example, supply-chain forecasts are typically made at a very granular level using very short time intervals (typically buckets of one week). In these circumstances, both the mean and the variance of changes in the signal (per period) will probably be low relative to the level of noise, thus the theoretical limit of forecast performance is likely to stay close to the ratio of 0.5. Lower ratios are possible for series with complex signal patterns, but these are liable to be more difficult to forecast than those with a simple signal. So we would not expect to see performance much better than this limit because the theoretical possibility of improving performance would be offset by the practical difficulty of achieving it. From a practical point of view, the proposed standards could be the best we can hope to achieve.

In summary, an unavoidability ratio of 0.5 in terms of MSE or 0.7 with respect to the MAE represents a useful estimate of the lower bound for forecast error in a range of circumstances. The upper bound is defined by the naïve forecast itself, so that a rational forecast process will normally produce a ratio between 0.5 and 1.0. The better the forecasting methodology, the closer the statistic will be to 0.5; in some circumstances it may be possible to better this. Potentially, then, this insight might provide a useful way of measuring forecast quality; the only way to assess quite how useful is through empirical work.

So much for the theory. What about the practice?

THE EMPIRICAL EVIDENCE

We carried out two tests comparing the performance of a set of forecasts against the respective naïve forecasts. For reasons of simplicity, absolute errors were used and compared to a theoretical lower bound of 0.7.

The first test (Unit A) used 124 product SKUs over 52 consecutive weekly buckets. The sample is from a fast-moving consumer-goods manufacturer whose business is characterised by a high level of promotional activity, and thus incorporates extensive manual intervention of statistical forecasts based on market intelligence. These are circumstances where it might be possible to significantly better the theoretical minimum. The distribution of errors relative to those from the naïve forecast is shown in Figure 1.

The second example (Unit B) comes from a consumer-durables business with a very fragmented product portfolio. There is a lesser degree of manual intervention in the (statistical) forecast process, but items with intermittent and lumpy demand are common. In this case, the sample comprised 880 SKUs across 28 consecutive monthly buckets. With monthly buckets, we might expect to see less noise and more change in the signal, thus making ratios below 0.7 more likely.

There are two striking things about these examples.

First, relatively few items have a ratio that falls below 0.7 (2% in the case of Unit A, 9% for Unit B), and almost none fall below 0.5. This suggests that a ratio of somewhere around 0.5 (even using the MAE, lower using the MSE) may well

Figure 1. The Unavoidability Ratio (Absolute Errors Relative to Those of a Naïve Forecast) for Unit A

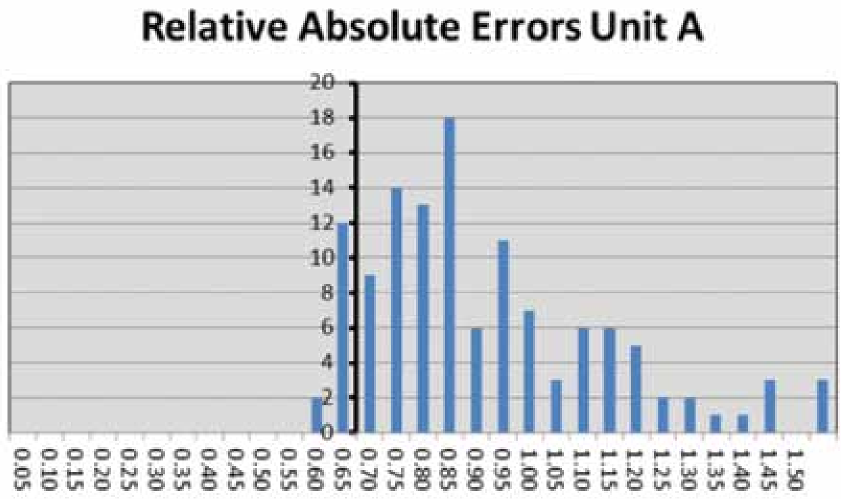
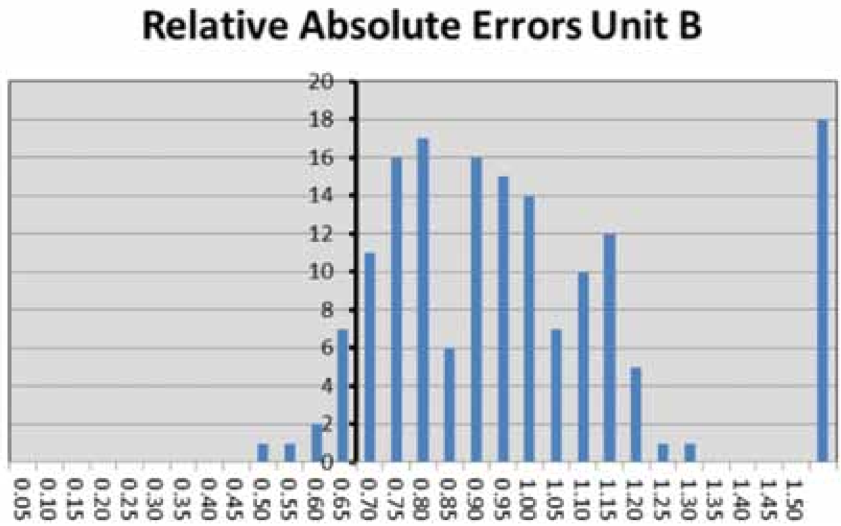


Figure 2. The Unavoidability Ratio (Absolute Errors Relative toThose of a Naïve Forecast) for Unit B



represent a useful "lower bound" benchmark in practice.

Note that products like Units A and B (high levels of manual intervention and intermittent demand patterns) challenge the robustness of the avoidability principle. Even here, the unavoidability ratio seems to provide a sound basis for estimating the performance potential that can be achieved by any forecast process, not only in principle but in practice. This result opens up the prospect of a wide range of practical applications including meaningful benchmarking and forecast-tracking techniques.

The second striking point is that both cases have a large number of SKUs with ratios in excess of 1.0 (27% for Unit A and 26% for Unit B), meaning that forecast performance was worse than the naïve, most likely the result of inappropriate manual interventions in the forecast process. Mike Gilliland (2013) considers this situation to be a case of negative Forecast Value Added (FVA). It certainly exposes significant potential for improvement in forecast quality; it also shows that while we may theoretically benefit from making intelligence-driven interventions in the forecasting process, these benefits are often not realised in practice, as pointed out by Goodwin and Fildes (2007).

Of course, more work is needed to validate and then build on the theoretical foundations established here. Crucially, more empirical work is needed to determine how robust the approach is in a wider range of less amenable forecasting situations, such as products with pronounced seasonal patterns (for example, daily sales data in a retail environment). There may also be ways in which any shortcomings in the approach can be mitigated in practice.

THE NEXT STEP

While absolute precision in benchmarking forecasting performance is some distance off – and may prove impossible – our evidence suggests that it is possible to set rational quality criteria with more confidence than hitherto thought possible. In turn, this could open the way to developing approaches to measuring and managing forecast performance that are more useful in practice than existing methodologies.

To operationalize these insights and assess their usefulness in practice, I would welcome participation from companies in a collaborative effort to further test the methodology and help develop and refine practical applications of this approach. Please contact me at the address below for further details.

REFERENCES

Armstong, S. & Collopy, F. (1992). Error Measures for Generalizing about Forecasting Methods: Empirical Comparisons," *International Journal of Forecasting*, 8, 69-80.

Boylan J. (2009). Towards a More Precise Definition of Forecastability, *Foresight*, Issue 13 (Spring 2009).

Catt, P. (2009). Forecastability: Insights from Physics, Graphical Decomposition, and Information Theory, *Foresight*, Issue 13 (Spring 2009).

Fildes, R. & Goodwin, P. (2007). Good and Bad Judgement in Forecasting: Lessons from Four Companies, *Foresight*, Issue 8 (Fall 2007).

Gilliland, M. (2013). FVA: A Reality Check on Forecasting Practices, *Foresight*, Issue 29 (Spring 2013), 14-18.

Goodwin, P. (2013). Theoretical Gains in Forecasting Accuracy Relative to Naïve Forecasts, Working paper, University of Bath.

Goodwin, P. (2009). Taking Stock: Assessing the True Cost of Forecast Errors, *Foresight*, Issue 15 (Fall 2009).

Hoover, J. (2009). How to Track Forecast Accuracy to Guide Forecast Improvement, *Foresight,* Issue 14 (Summer 2009).

Hoover, J. (2006). Measuring Forecast Accuracy: Omissions in Today's Forecasting Engines and Demand Planning Software, *Foresight*, Issue 4 (June 2006).

Hyndman, R. (2006). Another Look at Forecast Accuracy Metrics for Intermittent Demand, *Foresight*, Issue 4 (Summer 2009).

Pearson, R. (2010). An Expanded Prediction Realisation Diagram for Assessing Errors, *Foresight*, Special Issue: Forecast Accuracy Measurement: Pitfalls to Avoid and Practices to Adopt.

Schubert, S. (2010). Forecastability: A New Method for Benchmarking and Driving Improvement, *Foresight*, Issue 26 (Summer 2012).

Theil, H. (1966). *Applied Economic Forecasting*, Amsterdam: North-Holland.

Steve Morlidge is the author of *Foresight*'s multipart series *The Forecasting Process: Guiding Principles,* which presented a compelling argument "for forecasters to move beyond the exchange of experience and simplistic peddling of best practice to lay out a set of principles that collectively define forecasting craftsmanship." He is also coauthor of the book *Future Ready,* which draws lessons from his 30 years of experience designing and running performance-management systems.

steve.morlidge@satoripartners.co.uk

Forecast Quality in the Supply Chain

STEVE MORLIDGE

PREVIEW *Building on his two previous publications in* Foresight *on the measurement of forecastability, Steve shows how forecast quality can be objectively measured using the relative absolute error (RAE) metric and how this metric can be used to reveal the potential for improvements in forecast accuracy. He presents compelling evidence that many companies fail to achieve levels of relative error that are better than a simple "same as last period" naïve forecast, and that around 50% of individual forecasts fail to meet this benchmark. He makes it clear that, while there is a great need for improvement in forecast quality, there is the potential for forecasters to accomplish just such improvement.*

FORECAST QUALITY

Real-life experience suggests that forecast accuracy is a slightly slippery idea, one that is dependent on the context in which the error sits. For example, we would expect an "accurate" engineering blueprint to have much less error than a line drawing used for marketing purposes. Similarly, driving a golf ball 200 yards that lands within five yards of where the golfer was aiming might be called "accurate," whereas a putt that misses the hole by five yards certainly would not.

So it is with forecasting. We would expect a stable, "easy to forecast" data series to have lower errors than one that was more volatile. For that reason, in this article I use the word *quality* to measure forecast performance in a manner that reflects the forecastability of the data. Forecasting performance cannot be measured, or managed, in the absence of context.

THE STORY SO FAR

In the Summer 2013 issue of *Foresight*, I contributed to the long-running debate about forecastability (Morlidge, 2013). Here is the thrust of the arguments I made:

1. All extrapolative (time-series) methods are based on the assumption that the signal embedded in the data pattern will continue into the future. These methods thus seek to identify the signal and extrapolate it into the future.
2. Invariably, however, a signal is obscured by noise. A "perfect" forecast will match the signal 100% but, by definition, cannot forecast noise. So if we understand the nature of the relationship between the signal and noise in the past, we should be able to determine the limits of forecastability.
3. The most common *naïve* forecast uses the current period actual as the forecast of the next period. As such, the average forecast error from the naïve model captures the level of noise plus changes in the signal.
4. Thus the limit of forecastability can be expressed in terms of the ratio of the actual forecast error to the naïve forecast error. This ratio is generally termed a relative absolute error (RAE). I have also christened it the *avoidability ratio*, because it represents the portion of the noise in the data that is reduced by the forecasting method employed.
5. In the case of a *perfectly flat signal* – that is, *no trend or seasonality in the data* – the best forecast quality achievable is an RAE = 0.7. So unless the data have signals that can be captured, the best forecast accuracy achievable is a 30% reduction in noise from the naïve forecast.
6. An RAE =1.0 should represent the worst forecast quality standard, since it says that the method chosen performed less accurately than a naïve forecast. In this circumstance, it

might make sense to replace the method chosen with a naïve forecasting procedure.

The RAE has theoretical underpinnings and convincing empirical support. It provides us with a way of measuring the upper and lower bound of error and thus to make objective judgments about forecast quality. Specifically, it tells us whether a forecast has added value compared to the alternative "most primitive" method of forecasting (i.e., same as last period) and offers an obvious course of action if it is not possible to improve on the naïve.

Closely related to the RAE concept is that of *forecast value added* (Gilliland, 2013). However, FVA was created to accomplish a quite different goal: assessing process improvement, not relative error. It determines which steps in a process add value (improve accuracy), starting with a comparison against the naïve error, thus allowing distinctions between "acceptable" and "unacceptable" levels of accuracy. It is not prescriptive about what accuracy measures should be used, and consequently the concept of a lower bound for error (best accuracy potential) is lost.

I attempted to back up my assertions about forecast quality with a small-scale empirical test on corporate data. That test demonstrated that, while it is possible to improve upon an RAE of 0.7 as the result of the data having trend or seasonality or other signals, few forecasts did so. For those that did, an RAE of about 0.5 seemed to represent a practical limit on what could be achieved. This may be because while, in theory, a volatile signal makes it possible to deliver a lower RAE, in practice the more complex the signal the more difficult it is to forecast.

RESULTS FROM THE M3 COMPETITION

My second article (Morlidge, 2014) was an analysis of the data from the largest and most prestigious piece of academic empirical work on forecast accuracy: the *M3 forecasting competition*. While the M3 findings are highly relevant to the practical challenge of forecasting, most forecasting practitioners seem, unfortunately, to be unaware of it.

The M3 competition used 3,003 different data series – a combination of yearly, quarterly, and monthly – classified as "Micro," "Industry," "Macro," "Finance," "Demographic," and "Other." Each of these was forecast by experts using one of 24 different methods. I took the segment of the M3 data that is most relevant to supply-chain practitioners – the 334 short-run (one-month ahead) forecasts of industry data. For each forecast method, I calculated RAEs for each of these series and reported the weighted median RAE over all series. **Table 1** is taken from the earlier article.

Key Points

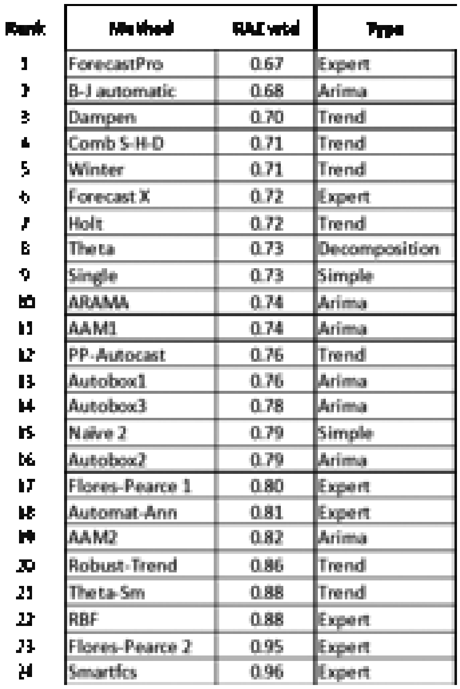
- Attempts to measure the performance of forecasting methods or processes have focused on forecast accuracy, using metrics based on forecast errors such as the MAPE. But these metrics give neither a true picture of forecasting performance nor of the potential for improvements in forecast accuracy. Practitioners should instead measure *forecast quality*, a measure of forecast performance that allows for forecastability — that is, *potential* forecast accuracy.

- The limit of forecastability can be expressed in terms of the ratio of the actual forecast error to the naïve forecast error. This ratio is generally termed a *relative absolute error* (RAE). Based on a sample of detailed demand forecasts from nine companies, we find that on average the RAE will rarely be below 0.7 and that RAEs of individual forecasts below 0.5 are an exceptional occurrence.

- Far too often, the forecast methods employed have an RAE in excess of 1.0 and so are less accurate than a *naïve* model that simply extends the current value to the future. Such a situation illustrates how important it is to choose the right forecasting methods and make wise adjustments where necessary – and how poorly this is so often done. It is also evidence of the significant scope for improvement in forecast performance.

- This evidence also suggests that investment in sophisticated forecasting software is, by itself, no guarantee of success. Without effective measurement practices and adequate education and training of forecasters, there is a strong possibility that such investments will generate little or no return.

Table 1. The Weighted Average RAE for One-Month Forecasts



Rank	Method	RAE wtd	Type
1	ForecastPro	0.67	Expert
2	B-J automatic	0.68	Arima
3	Dampen	0.70	Trend
4	Comb S-H-D	0.71	Trend
5	Winter	0.71	Trend
6	Forecast X	0.72	Expert
7	Holt	0.72	Trend
8	Theta	0.73	Decomposition
9	Single	0.73	Simple
10	ARAMA	0.74	Arima
11	AAM1	0.74	Arima
12	PP-Autocast	0.76	Trend
13	Autobox1	0.76	Arima
14	Autobox3	0.78	Arima
15	Naïve 2	0.79	Simple
16	Autobox2	0.79	Arima
17	Flores-Pearce 1	0.80	Expert
18	Automat-Ann	0.81	Expert
19	AAM2	0.82	Arima
20	Robust-Trend	0.86	Trend
21	Theta-Sm	0.88	Trend
22	RBF	0.88	Expert
23	Flores-Pearce 2	0.95	Expert
24	Smartfcs	0.96	Expert

The results validated most of the conclusions of the original M3 study. In particular, they demonstrated that statistically sophisticated or complex methods do not necessarily provide more accurate forecasts than simpler ones. Also, the relative ranking of the performance of different methods varies according to the accuracy measures being used, since the weighted RAE scores produced very different ranking than that for the median RAE. They showed an RAE of 0.7 represented "good" performance, and that around half of the methods achieved average scores close to this level.

Less expected, and perhaps more surprising, was the finding that all the 24 forecasting methods generated RAEs above 1.0 *more than 30% of the time*, which means that their performance was worse than that of a naïve forecast almost 1/3 of the time. While it is not unusual or unexpected for company forecasts to "fail," what was not expected was that forecasts produced by academic experts under controlled conditions (plenty of time, top-notch methods, no difficult-to-forecast intermittent demand/new products, etc.) would fail to this extent. These poor RAE scores were scattered throughout the 334 data series in the sample, meaning that these outcomes were not the result of specific data series being inherently difficult to forecast.

In summary, the average performance of any forecasting method may be less important than the distribution of its actual performance. At a practical level, however, we do not yet have the capability to identify the potential for poor forecasting *before the event*. It is therefore critical that actual forecast performance be routinely and rigorously measured after the event, and remedial action taken when it becomes clear that the level of performance is below expectations.

THE NEW STUDY

The purpose of this third article in the series is to submit the finding of the first two papers to more rigorous empirical challenge. I do this through the analysis of the actual forecasts from supply chain (product) data in a variety of industries.

The Goals

The suggestions from the evidence in our earlier studies are what we specifically aimed to test more rigorously now:

1. On average across the products in a company, an RAE will rarely be below 0.7. While it is theoretically possible to beat 0.7 when the pattern of signal is complex, in practice complex patterns are often more difficult to detect and forecast effectively than are simple patterns.

2. An RAE of individual product forecasts below 0.5 will be an exceptional occurrence. This seems to be the practical limit of what is achievable for any data series, regardless of the presence and type of signals.
3. At least 30% of individual forecasts will have an RAE above 1.0. In the absence of rigorous measurement using the RAE metric (and which is subsequently acted upon), at least 30% of data series will be forecast using inappropriate methods.

The Data

My samples of forecasts and actuals came from a variety of industries, covering at least 12 periods, which could be either weekly or monthly, and (in contrast to the forecasting competitions) represented low-level data, stock-keeping-unit (SKU) level or equivalent.

To make all the samples truly comparable and produce results that are most meaningful for supply chain professionals, the data should be sourced at the level and frequency at which the forecast is generated. For many businesses, forecasts are generated for SKUs by distribution location on a weekly basis. Forecast accuracy statistics can vary significantly depending on the level of granularity at which they are calculated due to the "netting off" effect on errors – whereby, for example, low-level errors of +10 and -10 combine to create a "perfect" forecast at higher levels. The RAE statistic is much less sensitive to this problem since both the numerator and denominator in the formula are subject to netting off, so the key conclusions arising from this study are not affected.

In all, I obtained nine anonymous samples, drawn from eight businesses operating in consumer (B2C) and industrial (B2B) markets. All these contributors used statistical forecasting packages. I analysed data for around 17,500 products for 29 periods on average, giving a third of a million data points in all. Most of the samples were provided in monthly buckets, and all but one used a forecast lag of one period ahead.

The Analysis

I calculated the RAE for every one of the 17,500 products. I then analyzed the distribution of individual product RAEs for each of the nine companies and calculated the median and weighted average RAE values companywide. I also calculated the traditional metrics of Forecast Accuracy (100% - MAPE), and Naïve Forecast Accuracy. (I had calculated these same statistics for the industry data from the M3 competition.)

THE NEW RESULTS

The results of the analysis are shown in **Table 2**.

1. The median RAE for each company was in the range 0.94 to 1.06 with an average of 1.02. Moreover, the 52% of the forecasts in our overall sample had RAEs above 1.0, with the best company result being 42% and the worst 62%. This result distressingly suggests that, on average, a company's product forecasts do not improve upon naïve projections. Keep in mind that these companies represent a wide spread of industries. Without further work, it is difficult to tell to what extent this is the result of poor model selection, inappropriate judgmental adjustments to the statistical forecast, or unforecastable data series. The demand-weighted average RAE was spread wider (from 0.89 to 1.89), but this was largely a result of two outliers. If these outliers were excluded, the range narrows to 0.81 to 1.14, and the average falls from 1.14 to 0.96 – very close to the average median RAE.
2. The RAEs for individual products rarely fall below 0.5. Only 5% of the 17,500 products had lower RAEs, and in only one company did more than 10% of the product forecasts achieve RAEs below 0.5. This adds credibility to our judgment that 0.5 represents a reasonable estimate of the practical lower limit for forecast error.

We can speculate on why 5% of the cases fell into this RAE<0.5 category. It may be that these products had a complex signal which was well forecast, but it could equally be the result of factors that have nothing to do with the quality of the statistical forecast process (e.g. products made to order), small sample sizes (since not all products were forecast all through the period), or simply that the forecasters were lucky.

3. There is essentially no correspondence between RAE and forecast accuracy (or MAPE), which means that the traditional forecast accuracy metrics give no guide to the product's degree of forecastability or its relative performance compared to a very crude benchmark. In **Figure 1**, the red line is fixed at an RAE of 1.0 and the blue and green lines mark the RAE thresholds of 0.5 and 0.7 respectively. The diamonds depict the median RAE (horizontal

Table 2. Forecast Accuracy and Quality Metrics by Company

	Forecast Characteristics				Distribution of Product Level Forecasts				Summary Statistics				
Name	Industry	Bucket	# Periods	# Products	RAE <0.5	RAE 0.5-0.7	RAE 0.7 -1.0	RAE >1.0	Median RAE	Wtd Av RAE	Forecast Acc	Naïve FA	Value Added
Business 1	B2C	Weekly	51	113	0%	0%	0%	0%	0.94	0.89	49%	44%	6%
Business 2	B2C	Weekly	30	484	0%	0%	0%	0%	1.15	1.04	77%	78%	-1%
Business 3	B2C	Monthly	12	0	0%	0%	0%	0%	0.97	0.81	34%	19%	15%
Business 4	B2B	Monthly	12	0	13%	0%	0%	0%	1.00	1.53	35%	58%	-22%
Business 5	B2B	Monthly	43	0	0%	0%	0%	0%	0.99	1.14	45%	52%	-7%
Business 6	B2B	Monthly	36	0	0%	0%	0%	0%	1.06	1.89	8%	52%	-43%
Business 7	B2B	Monthly	12	0	0%	0%	0%	0%	0.94	0.99	35%	34%	1%
Business 8	B2C	Monthly	34	0	0%	0%	0%	0%	1.05	0.87	53%	46%	7%
Business 9	B2C	Monthly	34	0	0%	0%	0%	0%	1.10	0.99	51%	51%	1%
Total/Average			29	597	1%	0%	0%	0%	1.02	1.13	43%	48%	-5%
Excluding Outliers										0.96			3%
M3 (av of 24 methods)	Various	Monthly	1	335	8%	13%	18%	61%	0.90	0.78	93%	91%	2%

Figure 1. Plot of Forecast Accuracy against RAE

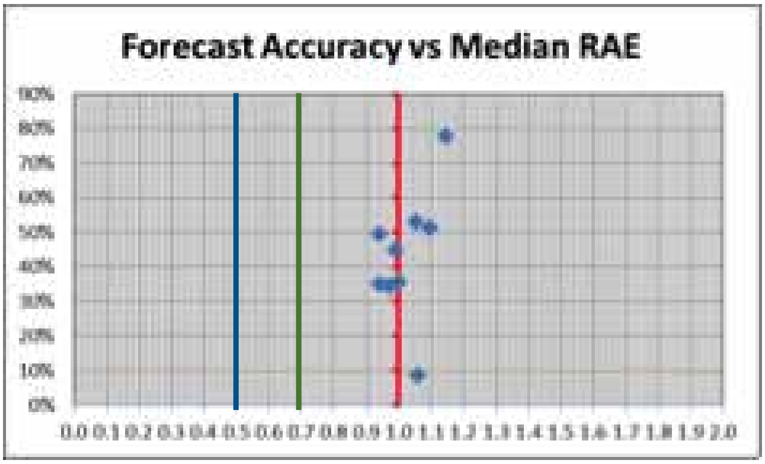


Figure 2a. Sample Companies

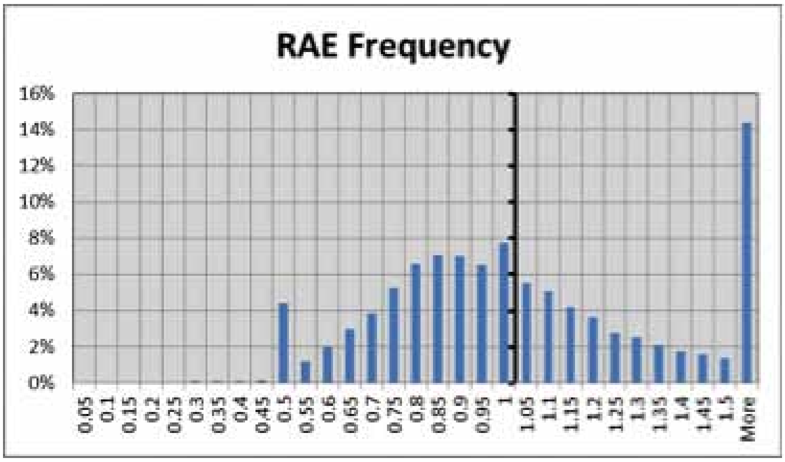
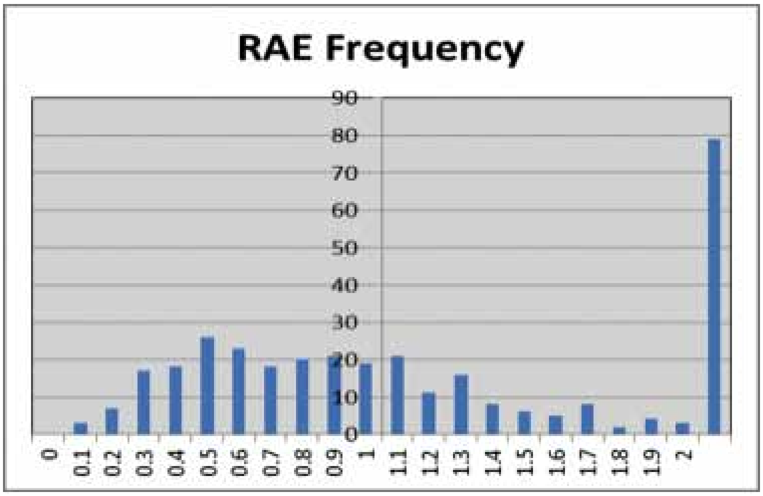


Figure 2b. M3 Industry Data



axis) vs. the Median Forecast Accuracy for the nine companies. No relationship is apparent between the two metrics. Indeed, the best as well as the worst Forecast Accuracy have similar RAEs.

COMPARISON WITH THE M3 COMPETITION

Figures 2a and **2b** compare the distribution of the RAE in our companies with that among the industry data in the M3 competition.

Although we selected that part of the data-set categories as "industry," the M3 data set is very different from the low-level supply chain data used in this study. The median Forecast Accuracy in the M3 is 93%, double that achieved by our sample companies.

Importantly, the naïve forecast accuracy of the M3 data is also double that of our sample. This means that the M3 data was much less volatile. Closer examination also reveals that the phenomena which cause supply chain forecasters so much trouble – products with intermittent demand and those at the beginning and end of product lives – are completely absent from the M3 data set. In our sample, at least a third of the actual values are zero, whereas there are none at all in the M3 data set.

So, the data used for the M3 appear to give an unrepresentative view of forecast accuracy and hence should not be used as a benchmark for evaluation of forecasting performance.

The absence of such "hard to forecast" data series may be the reason why the average RAE is so much lower (better) for the M3, but the incidence of high individual RAEs (above 1.0) is striking (over 50%), similar to that for our real-life data. This could be because, in the M3, forecasting methods were applied without discrimination as to whether they were appropriate for the individual data series.

M3 has a higher percentage of "good" forecasts (below 0.5); this, however, could be a reflection of the small sample size of the M3 data. We should not be surprised to see a high proportion of more extreme values.

CONCLUSIONS AND IMPLICATIONS

Even given the limited nature of the data, certain conclusions are clearly warranted, with implications that are significant for the practical task of forecasting in the supply chain.

1. An RAE of 0.5 is a good approximation to the lowest bound of forecast error: the best forecast that can be achieved in practice.
2. Traditional metrics such as MAPE are not helpful because they do not convey information about whether the forecast has the potential to be improved. By the same token, changes in the value of these accuracy metrics may represent not a change in the level of performance but rather a change in the volatility of the data series (as measured by the naïve forecast).
3. Many forecasting methods add little value, at least when they are applied inappropriately, and the performance of some is worse than one could achieve by adopting the prior period's actual as the forecast: a salutary (and likely sobering) thought for companies that have invested very heavily in software and the process around it.

To emphasize the operational implications:

1. In order to measure and manage the performance (quality) of any forecasting process, RAE should be routinely calculated at low levels of aggregation.
2. Given that RAE automatically adjusts for forecastability and its demonstrated usefulness in comparing the performance of forecast processes with very different characteristics, it should be valuable as a benchmark of forecast performance between products, geographies, companies, and industries.
3. Since the weighted average RAE is usually well above the lower bound of forecast error, there is significant scope for the improvement of forecast quality. And because roughly 50% of forecasts are above the upper bound of the RAE – all of which is avoidable – it is likely that the easiest way to make significant improvement is by eliminating poor forecasting rather than trying to optimise good forecasting.
4. In the supply chain, many hundreds or thousands of products need to be reforecast on a frequent basis. Again, rigorous measurement of forecast quality using RAE – ideally having split out the relative contribution of judgmental adjustment to the results – is indicated as the way to determine that the forecasts are being done effectively and to focus attention on those areas where they are not.
5. Investment in sophisticated forecasting software is, by itself, no guarantee of success. Without effective measurement practices and the education and training of forecasters, there is a strong possibility that such investment will generate little or no return.

NEXT STEPS

This work suggests that RAE provides forecasting practitioners with the ability to make well-founded judgments about the quality of forecasts and comparisons between forecasts in a way not hitherto possible. It also suggests ways in which this insight might be used in order to improve performance. For example, using the naïve forecast would improve all those forecasts with RAE in excess of 1.0. Quite rightly, this is not a strategy that any forecaster would be comfortable adopting unless all other ways of improving matters had been exhausted. It also does not help us to make improvements in those forecasts with RAE below 1.0.

The next article in this series will address this topic by exploring practical ways this approach can be used to drive improvements in forecast quality.

REFERENCES

Gilliland, M. (2013). FVA: A Reality Check for Forecasting Practices, *Foresight*, Issue 29 (Spring 2013), 14-19.

Morlidge, S. (2014). Do Forecasting Methods Reduce Avoidable Error? Evidence from Forecasting Competitions, *Foresight*, Issue 32 (Winter 2014), 34-39.

Morlidge, S. (2013). How Good Is a "Good" Forecast? Forecast Errors and Their Avoidability, *Foresight*, Issue 30 (Summer 2013), 5-11.

Steve Morlidge is author of *Foresight*'s multiple-part series on the Guiding Principles of the Forecasting Process (2011-2012). His analysis of the forecastability issue is an ongoing project. If you would like to participate or contribute data, please contact him at:

steve.morlidge@catchbull.com

Using Relative Error Metrics to Improve Forecast Quality in the Supply Chain

STEVE MORLIDGE

PREVIEW *How can we identify our best opportunities to improve forecast accuracy? Steve Morlidge concludes his four-part* Foresight *series on forecast quality by offering an approach based on (a) product volumes and variability, and (b) a forecastability metric that assesses forecast accuracy in relation to the accuracy of a naïve (i.e., no change) forecast. The metric helps supply-chain forecasters set meaningful targets for improvement, quantifies the scope for improvement, and tracks progress toward final goals.*

INTRODUCTION

"This is too wishy-washy. You will have to do something about this."

This was one amongst the many comments made by *Foresight* editors on receipt of my last article (Morlidge, 2014b). In it, I had detailed the results of the survey of nine sets of supply-chain forecasts drawn from eight businesses, comprising over 300,000 data points in total. I measured the performance of all these forecasts using a Relative Absolute Error (RAE) metric, where actual forecast error is compared to the simple "same as last period" naïve forecast error.

My purpose was to assess forecast quality in the supply chain by determining practical upper and lower bounds of forecast error – the lower bound representing the best accuracy that can be expected, the upper bound the worst that should be tolerated. My results – printed in the Spring 2014 issue of *Foresight* – showed that there were very few forecasts that had forecast errors more than 50% better than the naïve forecasts. Thus, for practical purposes, the lower bound of forecast error for the granular supply-chain data is an RAE of 0.5

But also, and somewhat shockingly, I found that approximately 50% of the forecast errors were worse than those from the naïve forecasts, with an RAE > 1.0, the logical upper bound of forecast error. This is not a healthy situation: in principle, it should be easy to beat the naïve forecast. Failure to do so means that the forecast process is adding no value to the business. It also begs a couple of key questions: "What is causing this?" and "What can be done about it?"

This was the issue that frustrated *Foresight* editors, and quite rightly so. Improving the craft of forecast measurement is laudable, but if nothing can be done with the results then we have won no more than a Pyrrhic victory. No approach to measuring the quality of forecasts can, in itself, improve accuracy; it is a challenge for any measurement scheme, not just for RAE.

Therefore, in this current article, I will offer specifics on how to use the forecast-quality metric (RAE) in conjunction with product volumes to target efforts to improve forecast quality in the supply chain.

Before starting out on this quest, let me reprise some relevant points from my previous articles and explain their relevance to the task of forecasting in the supply chain.

BACKGROUND

My motivation has been to discover the upper and lower bound – the worst and best

Key Points

- The Relative Absolute Error (RAE) metric compares forecast error to the "same as last period" *naïve forecast error*. An RAE >1 suggests that forecast error is actually worse than the naïve forecast error, an untenable situation. Unfortunately, we've found that such a result occurs all too frequently with supply-chain data – indeed, about the half the time.
- We found that an RAE < 0.5 is so rare that 0.5 can be considered a practical lower bound for forecast error. This means that a forecast method is performing at capacity if its errors reach 50% below those of the naïve forecasts. The challenge then for supply-chain forecasters is to drive RAE down as close to 0.5 as possible.
- Efforts to improve forecast quality should concentrate on "high-volume – high-RAE products." This task is abetted by our findings that a large proportion of the opportunity for error reduction is concentrated in a small proportion of the product portfolio.
- Use the coefficient of variation (COV) to identify data series likely to feel the impact of one-off events. These are series where relatively large judgemental interventions need to be made to improve statistical forecasts. Where the COV is relatively low, the strategy should be to refine the forecasting method, allowing judgemental adjustments to statistical forecasts only where the case for making a change is overwhelming.

levels – of forecast error and, in the process, produce a metric that can be used to make objective judgements about forecast quality.

The Upper Bound

The upper bound is easy to establish: there is no good reason why any set of forecasts should have larger errors on average than forecasts produced by the most primitive forecast conceivable – a naïve forecast that uses the prior period's actual as a forecast. This upper bound provides a benchmark against which forecast performance can be compared. A Relative Absolute Error (RAE) of below 1.0 means that the average level of absolute errors from a forecast is lower than that of the naïve forecast; above 1.0 means that it is worse. But for practitioners working in the supply chain, the naïve forecast is more than a convenient benchmark.

Forecasting demand, and replenishing stock based on the demand forecast, is only economically worthwhile if it is possible to improve upon the simple strategy of holding a fixed buffer (safety stock) and replenishing it to make good any withdrawals in the period. This simplistic replenishment strategy is arithmetically equivalent to using a naïve forecast (assuming no stock-outs), since the naïve forecast is one of no change from our current level.

Safety Stock and Forecasting Value

The safety stock needed to meet a given service level is determined by our forecast errors. If the RAE of our forecasts is 1.0, yielding the same error on average as a naïve forecast, the buffer set by the naïve errors is appropriate. If our forecast has an RAE below 1.0, however, it means that the business needs to hold less stock than that indicated by the naïve. This is how forecasting adds value to a supply chain: the greater the level of absolute errors below those of the naïve forecast, the less stock is needed and the more value is added. Put simply, forecasting is not an end in itself, it is a means to an end; the end being a more efficient way of managing inventory (Boylan and Syntetos, 2006).

In order to assess the potential of a forecast to add more value (how much improvement it is possible to make), we need to be able to identify the lower bound of forecast error.

The Lower Bound

My first article in this series on forecastability

Forecasting demand, and replenishing stock based on the demand forecast, is only economically worthwhile if it is possible to improve upon the simple strategy of holding a fixed buffer (safety stock) and replenishing it to make good any withdrawals in the period.

included a demonstration of how the lower bound of error could be determined theoretically (Morlidge, 2013). It showed that the lower bound of forecast error is a product of (a) the level of random noise in a data series compared to the change in the signal, and (b) the volatility of the change in a signal. In the case of a signal with no trend, the theoretical lower bound of error was close to 30% below the naïve forecast, irrespective of the level of noise: i.e., an RAE of 0.7.

Trends, seasonal movements, and other systematic changes in the signal could theoretically lower (improve) the RAE further, but it was my speculation that the more changeable the signal is, the more difficult it is to forecast. In practical terms, I argued that it would be difficult for any forecast to better an RAE of 0.5, a hypothesis that was supported by my empirical work on supply-chain forecasts (Morlidge, 2014b).

THE PRACTICAL CHALLENGE

If 0.5 is accepted as a practical lower bound, then error in excess of an RAE of 0.5 is avoidable, while error below an RAE of 0.5 is unachievable and hence unavoidable. In principle, then, supply-chain forecasters should seek to drive RAE down as close to 0.5 as possible. However, they need to be mindful of the likelihood of increased difficulty of making incremental improvements the closer they get to the lower bound. Moreover, the value that forecasting generates for the business is related to the absolute amount of avoidable error, which is determined mainly by the product volume to be forecast. Hence analysts should be guided by the RAE weighted by volume, which is more meaningful as a measure of forecast performance than the unweighted average RAE.

With the requirement to forecast hundreds and often thousands of items by week or month, the practical challenges that supply-chain forecasters face are formidable. Some of these items can be volatile or intermittent, and may be affected by marketplace activity. In these situations, standard time-series methods cannot be used without adjustments and embellishments. Judgemental adjustments to statistical forecasts are therefore common (Goodwin and Fildes, 2007), and these are frequently based on input from people who are not forecasting experts. Worse, they may be motivated by "silo" concerns and pure self-interest (for example, submitting forecasts that are below target to ensure meeting a quota). Finally, forecasting software typically offers a bewildering array of methods and parameters and "black box" automatic algorithm selection processes that (as demonstrated by other research) cannot always be relied upon to produce acceptable results, even in controlled conditions (Morlidge, 2014a).

Given the nature of these challenges, any approach to improving the quality of supply-chain forecasts must help practitioners:

1. Focus on those areas where the effort/reward ratio is most favourable;
2. Devise approaches that help identify the likely cause of problems and tailor strategies to solve them; and
3. Set realistic goals mindful of 1 and 2 above.

FOCUS THE EFFORTS

Portfolio classification methods, such as "ABC," have been used extensively in inventory management as a way of helping practitioners develop differentiated approaches to the management of a portfolio, and to focus their efforts in those areas where they will be best rewarded (Synetos and colleagues, 2011).

Figure 1. RAE vs. Volume

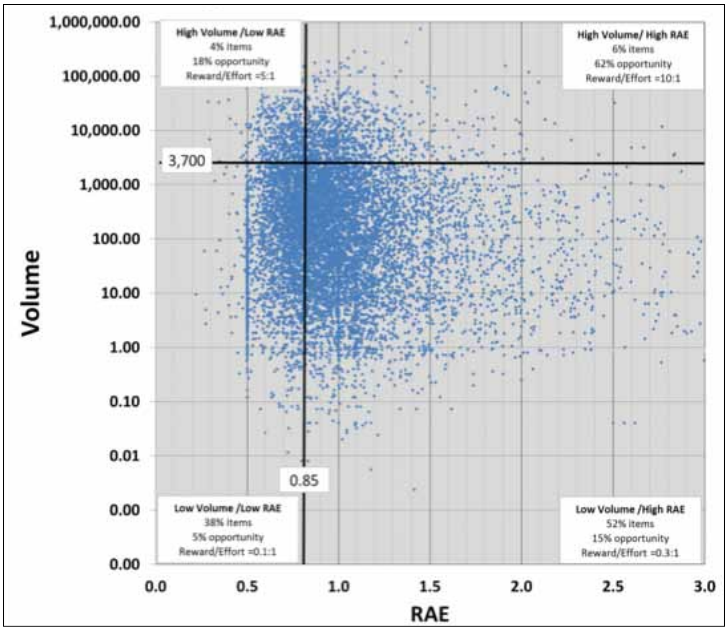
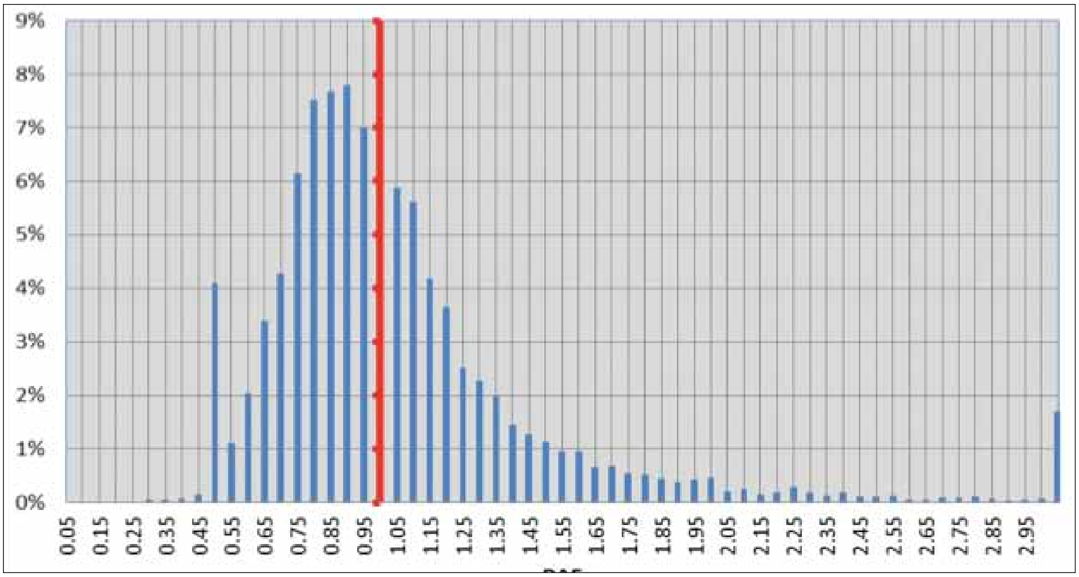


Figure 2. Distribution of RAE



One obvious way in which this approach could be applied to the challenge of forecast improvement is in helping practitioners target their efforts on those items with, at once, the poorest forecast performance (as measured by RAE weighted by volumes) and largest volumes.

This task will be easier if: (1) a large proportion of the opportunity (total amount of avoidable error in excess of 0.5 RAE) is concentrated in a small proportion of the product portfolio (true for our supply-chain data: approximately 20% of items contributed 80% of the avoidable error); and (2) forecast quality (RAE) is not strongly correlated with volume, as such a correlation might suggest that small-volume items are more difficult to forecast. In practice, we found this was not often the case, as large-volume products often did not have significantly lower RAE than low-volume products.

The "High Volume/High RAE" quadrant holds only 6% of the items but accounts for 64% of the opportunity, giving a very favourable effort/reward ratio. In this way, the focus of work to improve forecasting can be directed to those items where the greatest opportunities lie.

The first condition is the most important. A significant proportion of the opportunity (total amount of avoidable error) is typically concentrated in a small proportion of the product portfolio. For example, consider my previously used data comprising 11,000 items forecast in monthly buckets over a two-year period. **Figure 1** plots these 11,000 items (each represented by a dot) on a chart where the Y axis shows the average volume and the X axis marks forecast quality (RAE). (The volume axis uses a logarithmic scale so that the wide range of values can be displayed clearly, and so that any correlation between RAE and volume would be very obvious.) It is clear that no significant correlation exists in this case.

The histogram below the chart, **Figure 2**, shows a large number of items with RAE in excess of 1.0 (about 40%), all of which could be avoided by using the naïve forecast (although in practice this should be the last resort) and very few below with RAE 0.5.

I have drawn separators in Figure 1 to distinguish four quadrants. This shows that 80% of the avoidable error (opportunity) comes from items associated with the high RAEs of 0.85 or above.

Accuracy improvement here should be relatively easy to achieve. Further, 80% of avoidable error is with the largest (by volume)

20% of the items. As a result, the "High Volume/High RAE" quadrant holds only 6% of the items but accounts for 64% of the opportunity, giving a very favourable effort/reward ratio. In this way, the focus of work to improve forecasting can be directed to those items where the greatest opportunities lie.

This leads to the next question: how do we identify the best approach for exploiting these opportunities?

DEVISE IMPROVEMENT STRATEGIES

There are two ways that forecast quality can be improved:

1. Choosing better forecasting methods; and
2. Making better judgemental adjustments.

This is a truism that applies to all items, but the trick is to match the improvement strategy with the right part of the portfolio.

The approach outlined here involves isolating those parts of the portfolio where, in principle, judgement can make a significant contribution to forecast quality, and then taking steps to ensure that such judgement is used judiciously. Outside this zone, the use of judgemental adjustments should be restricted; instead, effort must be focused on optimising forecasting methods.

Figure 3 plots all the items in our sample portfolio on a second grid, which will help us select the most appropriate strategy to employ. This matrix is similar to the so-called ABC/XYZ approach used in the supply chain to help select the most appropriate replenishment and inventory policies.

As with the first classification grid, the Y axis represents volume and the horizontal line segregates the 20% of items that account for 80% of the avoidable error. However, here the X axis records the Coefficient of Variation (COV) of demand, which measures the volatility of the demand pattern. (I have calculated the COV as the ratio of the mean absolute deviation – rather than standard deviation – to the arithmetic mean, a calculation that mitigates the impact of more extreme observations.)

This approach is based on the reasonable assumption that, all things being equal, the less volatile the demand pattern (the lower the COV), the easier it will be for forecasting methods to successfully pick up and forecast the signal in the data.

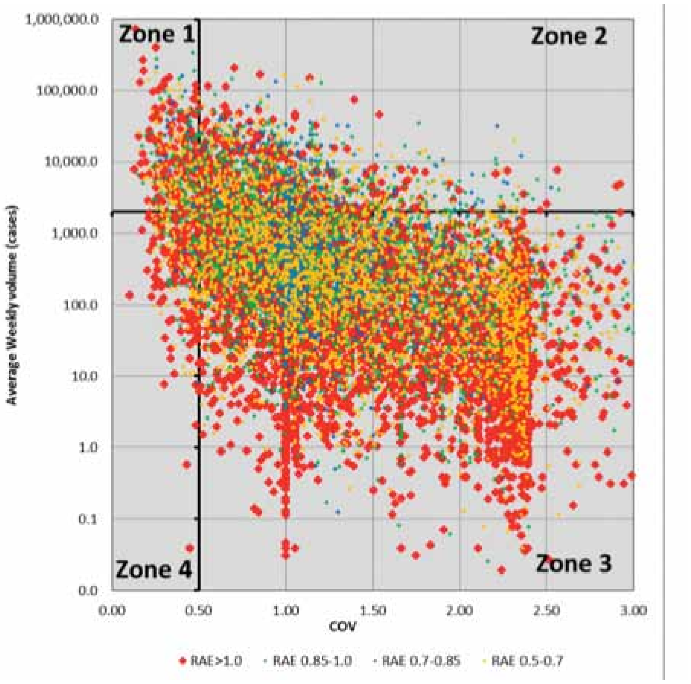
With lower COVs, there is less chance that judgemental intervention will improve forecast quality. On the other hand, higher COVs are more likely to be associated with data series heavily affected by sporadic events where relatively large judgemental interventions may be needed to improve statistical forecasts (Goodwin and Fildes, 2007).

The items are colour-coded based on their RAE:

RAE >1.0	= red
RAE 0.85 to 1.0	= amber
RAE 0.7 to 0.85	= green
RAE <0.7	= blue

A cursory visual inspection of the chart suggests that there is considerable scope for improvement, based on the widespread scattering of red items. To maximise the opportunities for meaningful improvement,

Figure 3. Volume vs. Volatility (COV) of Forecast Items [Color codes distinguish forecast quality (RAE)]



we must proceed in a structured, stepwise manner. This is my approach:

Priority 1: High-Volume/High-RAE Items

This is the part of the portfolio where the effort/reward ratio is most favourable, in that 6% of the items contribute 62% of the avoidable error. In Figure 3, these are (a) above the line and (b) coded with amber or red dots.

Some of these items are in Zone 1, where the COV is relatively low. For these the strategy should be to focus on refining the forecasting method (how data is cleansed, models selected, forecasts tracked), allowing judgemental adjustments to statistical forecasts only where the case for making a change is overwhelmingly favourable and the impact is likely to be significant (Goodwin and Fildes, 2007).

Zone 2 contains those items with a more volatile data pattern. Optimising the forecasting method here is more difficult given the volatile nature of the data series and impact of one-off events. The focus in this zone should be on the effective use of judgement. The exception to this may be items with a well-defined seasonal pattern, which could be forecast statistically without manual intervention despite having a high COV.

Zone 2 is the part of the portfolio where consensus forecasting techniques (statistical plus judgemental) are likely to add most value. That these items encompass a small proportion of the total number of items means that valuable management time can be focused very effectively. The success of these interventions can be quantified by measuring RAE before and after the consensus process, and using the forecast value added concept for the comparison (Gilliland, 2013). Since poor judgement is often manifest in consistent over- or underforecasting, managers should continuously monitor for bias.

Priority 2: High-Volume/Low-RAE Items

This second most interesting part of the portfolio comprised an additional 18% of the avoidable error. These items lie above the line and are colour-coded green or blue. For the green items, I'd recommend the same approach followed for Priority 1; that is, improving the statistical forecasts while discouraging the application of judgement except for those items with a high COV. Of course, it would not be worthwhile to work on the blue items, since they already have the very lowest RAE (lower than 0.7).

Priority 3: Low-Volume Items

In our sample, Zones 3 and 4 of the portfolio contain 90% of the items but only 20% of the avoidable error. Irrespective of the level of variation in the data series, they are unlikely to reward any efforts involved in a consensus forecasting process.

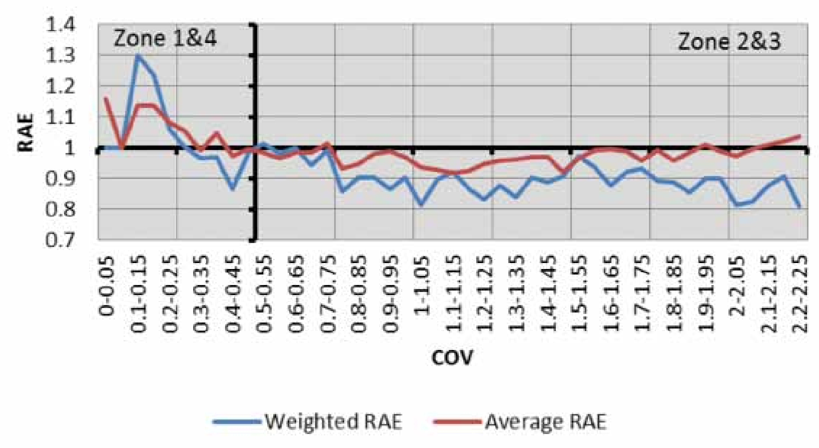
Instead, the focus should be using a very simple and conservative forecasting method, such as simple exponential smoothing (SES). The intermittent-demand items, which are most likely to be in Zone 3, should be forecast using SES or a variant of Croston's method (Synetos and colleagues, 2011). In some cases, where a data series approximates a random walk, the naïve model itself may be the best we can do. Perhaps these are not worth forecasting at all, using instead simple replenishment strategies or make-to-order (Boylan and Syntetos, 2006).

SETTING REALISTIC TARGETS

Because the portfolio analysis is an exercise that will be carried out only periodically, it will be necessary to continuously track forecast quality (Hoover, 2009) to check that the hoped-for results are delivered and to identify when performance levels start to drop, necessitating another review. The key question, however, is "What level of performance should we be aiming to achieve?" Clearly an RAE above 1.0 always flags a problem, and should be investigated (particularly if it is associated with a high-volume item), but what target should we be shooting for?

In a previous issue of *Foresight*, Sean Schubert suggests an approach based on the forecastability DNA of a product (Schubert, 2012), which takes account of factors other than the naïve forecast error. Here I propose adopting a similar approach by taking into account the volatility of the data series.

Figure 4. Average RAE Volatility (COV)



We have established that an RAE of 0.5 represents a practical lower limit of error in most cases. It would not be productive to adopt 0.5 as a target for small-volume items since the effort involved here probably could not be justified. For larger-volume items, Paul Goodwin has suggested a formula for setting sensible targets.

Goodwin's formulation is based on the assumption that the lowest RAEs are associated with the items with the most volatile signals, which are likely to be items with the highest COV. This is counterintuitive: COV is often considered to be a measurement of forecastability, with higher COVs indicating more volatility and thus greater difficulty in achieving any given level of forecast accuracy. But, as shown in **Figure 4**, as the COV increases, the weighted RAE tends to decline. Hence our argument is that we should set more stringent RAE targets for the higher COV items.

The logic underpinning this argument is this: if the product is unforecastable – if the naïve forecast error is totally driven by noise – an RAE below 1.0 is unachievable. If there is a signal in the data (trend, seasonal, external factor) then the product is potentially forecastable, and the RAE should be expected to be better (lower) than 1.0. And we see here that lower COV forecasts often perform very badly compared to the naïve, resulting in high RAEs.

Figure 4 plots the average and weighted-average RAE against COV for our sample. Figure 4 shows an increasing gap between the simple and weighted-average RAEs, reflecting that high-volume items/ high-COV items (i.e., those in Zone 2) have lower RAEs than those items with lower volumes.

Targets for High-COV Items

Figure 4 results suggest that the target for Zone 2 items (high volume, high volatility) should be a relatively low RAE, while the target for items in Zone 3 (low volume, high volatility) should be less ambitious on the grounds that we quickly reach diminishing returns.

Targets for Low-COV Items

In Zones 1 and 4 of Figure 3, which comprise items with low COV, our intuition is to expect lower levels of forecast error than in Zones 2 and 3 — that is, better RAE scores. Figure 4, however, shows that the lower the COV, the worse the RAE (in this case, the RAE is significantly higher than 1.0). Also, there is no consistent difference between the simple and weighted-average RAEs, meaning that high-volume items have been forecast no better than low-volume items on average. What is causing this pattern is not clear – it may be the result of poorly judged manual interventions or overfitting of forecasting models – but whatever the cause, it is clearly unacceptable, and reasonable to expect better RAE scores for items in Zones 1 and 4 (though perhaps not as high as in Zone 2).

In summary, items in Zone 2 should have the most stretching targets since this is where

Table 1. Performance Targets and the Scale of Potential Improvement

	Percentage of Items	Current RAE	Target Range	Improvement Potential
Zone 1	3%	1.01	0.7-0.85	14%
Zone 2	13%	0.92	0.5-0.7	25%
Zone 3	7%	1.05	>1.0	0%
Zone 4	77%	0.99	0.7-0.85	17%
Average	**100%**	**0.97**	**0.70**	55%

the greatest scope exists to add value by manual intervention, and Zone 3 the least stretching because the low volumes make it unrewarding to expend the effort required to achieve good RAE scores. The targets for Zones 1 and 4 lie in between these extremes, but should be achievable with less effort because judgemental intervention is less likely to be needed.

Based on this analysis, I have proposed targets for items in each of these four zones in our sample, compared to the historic performance (**Table 1**). The scale of potential improvements is very significant: avoidable forecast error (as measured by the weighted RAE) might perhaps be halved, with 71% of the total potential being contributed by 16% of the product portfolio. For the remaining 84% of items, the biggest contribution of this approach probably lies with the scope it gives to significantly reduce the amount of time and effort applied to forecasting them.

CONCLUSION

While it is unwise to make big claims based on one example, using RAE in conjunction with a small number of other easily calculated measures does appear to provide an objective and rational platform for constructing a set of forecast-improvement strategies tailored to a product portfolio. The goal is to maximize the overall benefit for a given level of effort.

Compared to a similar classification but based on conventional error metrics, RAE brings a number of benefits:

It identifies where the greatest opportunities lie by quantifying the scope for improvement and where it is concentrated in the portfolio.

It provides a quick and simple approach for dealing with items that are forecast poorly, and where the scope for improvement does not warrant the effort (the naïve forecast).

It helps set meaningful goals, tailored to the nature of the product and the role it plays within a portfolio. These can be used to quantify the scope for improvement and track progress.

Steve Morlidge is author of *Foresight's* multiple-part series on the Guiding Principles of the Forecasting Process (2011-2012). His analysis of the forecastability issue is an ongoing project. If you would like to participate or contribute data, please contact him at:

steve.morlidge@catchbull.com

REFERENCES

Boylan, J. & Syntetos, A. (2006). Accuracy and Accuracy Implications for Intermittent Demand, *Foresight*, Issue 4, 39-42.

Gilliland, M. (2013). FVA: A Reality Check on Forecasting Practices, *Foresight*, Issue 29 (Spring 2013), 14-19.

Goodwin, P. & Fildes, R. (2007). Good and Bad Judgment in Forecasting: Lessons from Four Companies, *Foresight*, Issue 8 (Fall 2007), 5-10

Hoover, J. (2009). How to Track Forecast Accuracy to Guide Forecast Process Improvement, *Foresight*, Issue 14 (Summer 2009), 17-23.

Morlidge, S. (2014a). Do Forecasting Methods Reduce Avoidable Error? Evidence from Forecasting Competitions, *Foresight*, Issue 32 (Winter 2014), 34-39.

Morlidge, S. (2014b). Forecastability and Forecast Quality in the Supply Chain, *Foresight*, Issue 33, (Spring 2014), 26-31.

Morlidge, S. (2013). How Good Is a "Good" Forecast? Forecast Errors and Their Avoidability, *Foresight*, Issue 30 (Summer 2013), 5-11.

Schubert, S. (2012). Forecastability: A New Method for Benchmarking and Driving Improvement, *Foresight*, Issue 26 (Summer 2012), 5-13.

Syntetos, A., Boylan, J., & Teutner, R. (2011). Classification of Forecasting and Inventory, *Foresight*, Issue 20 (Winter 2011), 12-17.

Measuring the Quality of Intermittent-Demand Forecasts: It's Worse than We've Thought!

STEVE MORLIDGE

PREVIEW *In this eye-opening article, Steve Morlidge shows that when our demand histories are intermittent, we should rethink the use of our most common accuracy metrics for selecting a best forecast method. The problem is acute because many software applications use these metrics for performance evaluation and method selection; in doing so, they potentially provide us with poor feedback and inferior models, resulting in harmful consequences for inventory management.*

INTRODUCTION

In most businesses, there are products that do not register a sale in every period, a fact that complicates the lives of forecasters. Many practitioners are aware that intermittent demand needs to be forecast in a different way from normal demand, using methods like Croston's. (Note: for a tutorial introduction to the forecasting of intermittent demand, see John Boylan's 2005 article in the premiere issue of *Foresight*.)

Indeed, forecasters often realize it is tricky to apply conventional forecasting metrics like MAPE (mean absolute percentage error) in this area, because of the small or zero denominator in the equation. But few will be aware that the problem goes deeper than this: conventional accuracy metrics like MAD (mean absolute deviation) and MAPE can give misleading signals about forecasting performance and steer us to select poor models; this has potentially severe implications for inventory management, where forecasts are used to drive replenishment orders in a supply chain.

THE PROBLEMS WITH INTERMITTENT DEMAND

Intermittent demand has always presented problems for forecasters.

The main difficulty arises because the data that forecasters rely upon to make predictions is sparse: periods with positive values are often separated by a number of periods with zero values. As a result, it is difficult to identify trends and other patterns. And because it is also difficult to estimate which periods in the future will register some activity and which will be empty, most forecasters don't even try; instead, they seek to forecast an average (mean) value over time.

Many businesses deal almost exclusively in products that exhibit intermittent patterns of demand, and even those with more consistent demand patterns will encounter this problem if the choice is made to use narrower time buckets (e.g. weekly or even daily) for forecasting.

The difficulty in forecasting intermittent demand is compounded by the problem of measuring the quality of the results. It has long been recognised that intermittent levels of demand undermine the usefulness of traditional forecast error metrics, like MAPE. Because the denominator in the MAPE is the actual demand, a zero denominator will yield an infinite value for this metric. This is the denominator problem. In his 2006 article in *Foresight*, Jim Hoover describes just how poorly software solutions deal with the problem, some of which exclude the periods of zero actual demand for the MAPE calculation.

Key Points

- Measuring the forecasting performance of intermittent-demand models is a much bigger problem than most of us imagine. Intermittent levels of demand undermine the usefulness of traditional forecast error metrics like MAPE because a zero denominator (for periods when there is no demand) will yield an infinite value for the MAPE. This is the denominator problem.

- And there is an even bigger problem: the numerator problem. Instead of guiding us to the mean of a distribution, a metric based on absolute error – such as mean absolute deviation (MAD) – guides us to a model that predicts the median level of demand, which is the more common value in the intermittent-demand series. So, if 50% or more of the values are zero, the median and hence the "optimum" forecast will be a forecast of zero, irrespective of the size of the nonzero values.

- Various solutions to the numerator and denominator problems are discussed and evaluated. My solution to both problems is to separately measure the two components of forecast error – bias and magnitude of error – and then appropriately combine them. I call the resulting metric the bias-adjusted error metric.

Suggestions to address this problem include:

- **A Denominator-Adjusted MAPE** (DAM), in which each period of zero demand is represented by a 1, as if one unit had been demanded (Hoover, 2006). Still, with a small quantity in the denominator, small absolute errors can translate into extremely large percentage errors, exploding the MAPE, thus giving a distorted picture of forecast accuracy.
- Substituting the **MAD/MEAN** for the MAPE (Kolassa & Schutz, 2010). The two are similar in interpretation: while the MAPE is the mean of the absolute percentage errors, the ratio MAD/MEAN is the mean absolute error as a percentage of mean demand. However, in the MAD/MEAN, periods of zero actual demand are averaged in the denominator with the positive demands of other periods, avoiding the exploding MAPE.
- Using the **Mean Absolute Scaled Error (MASE)** in lieu of the MAPE (Hyndman, 2006). The MASE differs from the MAPE in that it calculates the forecast errors made as a percent of the in-sample (rather than forecast) errors from a naive model. It is similar to the MAD/MEAN in that both use the MAD in the numerator. The denominator elements of the MAD/MEAN, however, are the actual demands and not the errors from the naïve model.
- The **Relative Absolute Error** metric (Morlidge, 2013) is similar to the MASE, but uses the naïve error as the denominator from the same range of periods as the numerator – the range, as applied, being the out-of-sample (i.e., the forecast) periods.

All of these MAPE alternatives work by changing the denominator so that zeros do not explode the metric.

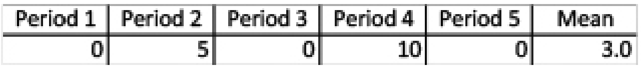
And there is an even bigger problem than this, one that has largely escaped the notice of practitioners and researchers: the numerator problem.

THE NUMERATOR PROBLEM

To understand the numerator problem, consider this simple example of an intermittent-demand series.

Take the sequence of demands shown in **Table 1**.

Table 1. An Example of Intermittent Demand



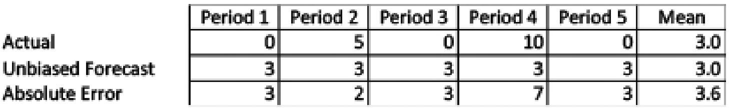
Period 1	Period 2	Period 3	Period 4	Period 5	Mean
0	5	0	10	0	3.0

What is the best forecast for this sequence?

On average, it would be best to forecast the mean=3 for each period, since total demand over the 5 periods is 15 units. And, as shown in **Table 2**, the average absolute error for this forecast is 3.6.

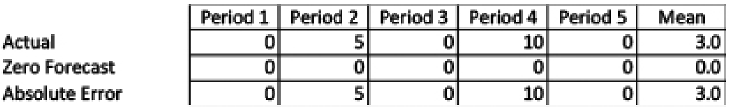
But look what happens if we make what seems to be an unthinkable forecast: zero for each period, an example proposed by Teunter and Duncan (2009). As shown in **Table 3**, the average absolute error is now just 3.0!

Table 2. Errors Associated with a "Perfect" Forecast



	Period 1	Period 2	Period 3	Period 4	Period 5	Mean
Actual	0	5	0	10	0	3.0
Unbiased Forecast	3	3	3	3	3	3.0
Absolute Error	3	2	3	7	3	3.6

Table 3. Errors Associated with a Zero Forecast



	Period 1	Period 2	Period 3	Period 4	Period 5	Mean
Actual	0	5	0	10	0	3.0
Zero Forecast	0	0	0	0	0	0.0
Absolute Error	0	5	0	10	0	3.0

So it appears that the zero forecast is better than that which correctly forecasts the mean demand of 3.0. This would be true regardless of how large the nonzero demands were in periods 2 and 5. How can this be?

The reason we get these apparently nonsensical results is because of a fundamental misconception: most of us probably assume that the average absolute forecast error metric (MAD) will guide us to select the best forecast method, the one that gives us a forecast closest to the mean demand pattern. But alas, this is not the case: instead of guiding us to the mean of a distribution, it guides us to the median, which is the most common value in the intermittent-demand series. If more than half of all periods exhibit zero demand, then the median will be zero.

So the average absolute error metric optimises on the median – not the mean – of the probability distribution. The mean and the median are the same if the probability distribution is symmetric – like the normal distribution – but not if the distribution is skewed, as is the case with intermittent-demand series: if 50% or more of the values are zero, the "optimum forecast" will be a forecast of zero, irrespective of the size of the nonzero values.

As you would suppose, the consequences of the numerator problem can be significant.

The main implication for forecasting practitioners is that it means we cannot judge how good our intermittent-demand forecasts actually are by using metrics like the MAD or MAPE. And it also means that we cannot rely on forecast algorithm selection methods that use the absolute error, when it comes to selecting the best forecast model.

Given this problem, one that is well known to statisticians (Hanley et al., 2001; Syntetos & Boylan, 2005), it will probably surprise practitioners to discover that the majority of academic research into different methods for forecasting intermittent demand – where the consequences are most acute – uses absolute error measures to analyse the results. Indeed, it has recently been suggested that this may be the reason why there has been so little consistency in the findings of research in this area (Teuntner & Duncan, 2009).

SOLUTIONS TO THE NUMERATOR PROBLEM

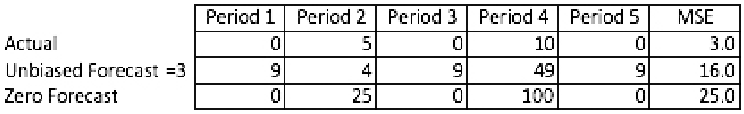
Since no business that has a significant number of products displaying intermittent demand can ignore the problem, what are the solutions?

A good solution should generate a forecast that optimises on the **mean demand** – not median demand. At a practical level, it is also important that the chosen metric is simple to calculate and easy to understand and explain. It should also work for ordinary (non-intermittent) demand since it is impractical to have different metrics for the two classes of demand, particularly since the dividing line between them is not easy to define.

Use the Mean Squared Error (MSE)

One option is to compare methods using mean squared error instead of the mean absolute error. As shown in **Table 4**, use of the MSE for the intermittent series in Table 1 would have correctly selected the best forecast of mean demand (3.0) rather than the median of 0. The MSE for this unbiased forecast is 16.0, while that for the zero forecast is 25.0.

Table 4. Squared Errors for the Forecasts in Tables 2 and 3



	Period 1	Period 2	Period 3	Period 4	Period 5	MSE
Actual	0	5	0	10	0	3.0
Unbiased Forecast =3	9	4	9	49	9	16.0
Zero Forecast	0	25	0	100	0	25.0

While this metric correctly finds the unbiased forecast at the mean (3) to be better than the zero forecast at the median (0), it comes with a major concern. Because of the squaring of errors, the MSE gives great if not extreme weight to "faraway" errors, with the potential to create a distorted impression of the impact of forecast error on the business (excessive safety stock). This is a particular problem for intermittent-demand series, which are by definition more volatile than "normal" data series and carry greater risk of outliers.

Direct Measurement of Inventory Costs and Service Levels

Another option involves measuring the impact of error on inventory or service levels directly (Teutner & Duncan, 2009; Wallstrom & Segerstedt, 2010). Doing so, however, is complicated and problematic since the relationship between error and the business impact will vary from product to product.

For example, the business impact of overforecasting will be very high if the product is perishable (e.g. fresh salads) or the cost of production is high (e.g. personal computers). In these circumstances, the impact of forecast error on stock levels is the primary concern. If the margin on a product is high or it is not perishable, and there is a risk of losing sales to competition, then the business is likely to be very sensitive to underforecasting (e.g. ice cream). Here, the impact of error on service levels is the most significant factor.

As a result, to measure the business impact of forecast error directly in a satisfactory manner, one needs a way of recognising those product characteristics that matter. It would be desirable to find a single metric that enables us to strike a balance between different types of impact – for example, the trade-off between the cost of higher stocks with the benefits of having a better service level.

Lastly, while it is easy enough to add up error to arrive at a measure of forecast quality for a group of products, it is less easy to do the same for a metric such as service level, particularly if different products have different target service levels.

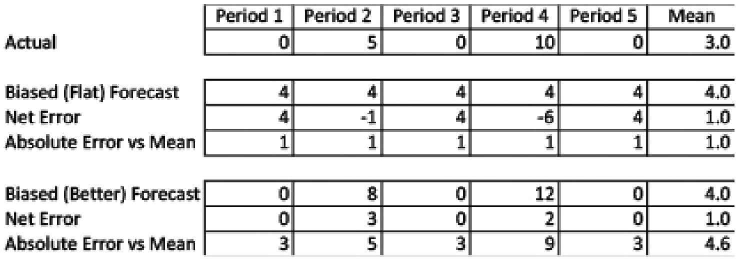
Mean-Based Error Metric

Some authorities (Wallstrom & Segerstedt, 2010; Kourentzes, 2014; Prestwich and colleagues, 2014) have proposed calculating forecast errors by comparing a forecast with the series mean over a range of actual values rather than the actual for each period.

This has the merit of simplicity and solves the denominator problem (unless every period demand is zero). However, while it successfully captures how well a forecast reflects the actual values on average – that is, it effectively measures bias – it ignores how far adrift the forecast is on a period-by-period basis. In effect, it assumes that all deviations from the mean demand represent noise.

This view can lead us astray when forecasts are equally biased, as the highly simplified example in **Table 5** demonstrates.

Table 5. Comparing Forecasts to the Mean Can Create a Misleading Impression of Forecast Performance



	Period 1	Period 2	Period 3	Period 4	Period 5	Mean
Actual	0	5	0	10	0	3.0
Biased (Flat) Forecast	4	4	4	4	4	4.0
Net Error	4	-1	4	-6	4	1.0
Absolute Error vs Mean	1	1	1	1	1	1.0
Biased (Better) Forecast	0	8	0	12	0	4.0
Net Error	0	3	0	2	0	1.0
Absolute Error vs Mean	3	5	3	9	3	4.6

Both forecasts are similarly biased over the range (both overforecast by an average of 1). Using this mean-based metric, however, the flat forecasts (=4) look significantly better because they are consistently close to the period average. On the other hand, the bottom set of forecasts looks mediocre (the absolute error against the mean being 4.6 compared to 1 for the first forecast) despite better capturing the period-by-period change in the demand pattern. The relative superiority of this bottom set of forecasts can be demonstrated without working through the detailed safety stock calculations: in the case of the flat forecasts, additional safety stock would need to be held to avoid stock-outs in periods that were underforecast (periods 2 and 4).

THE BIAS-ADJUSTED ERROR

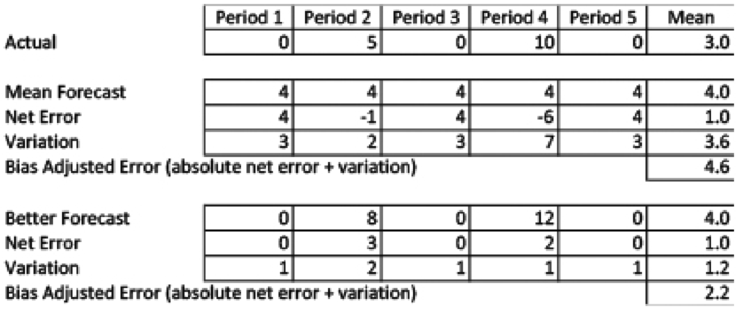
The approach I propose involves separately measuring the two components of forecast error – bias and dispersion of error – and then appropriately combining them. Minimizing bias is important because it ensures that, over time, we will not have consistently too much or too little stock on hand to meet demand. Dispersion of error has a direct impact on the safety stock needed to meet service-level targets.

In contrast, conventional metrics lump together bias and dispersion because they measure variation of the errors from zero, rather than from the mean of the errors. It can be enlightening to distinguish and separately report these two components.

- First, calculate bias by the mean net error (MNE).
- Second, calculate the magnitude of variation of error around the MNE.
- Finally, add the MNE (expressed in absolute terms) and dispersion measurement.

Table 6 illustrates the calculations. The appendix has a fuller explanation of the calculation method.

Table 6. The Bias-Adjusted Error Metric Correctly Reflects Both Bias and Variation



	Period 1	Period 2	Period 3	Period 4	Period 5	Mean
Actual	0	5	0	10	0	3.0
Mean Forecast	4	4	4	4	4	4.0
Net Error	4	-1	4	-6	4	1.0
Variation	3	2	3	7	3	3.6
Bias Adjusted Error (absolute net error + variation)						4.6
Better Forecast	0	8	0	12	0	4.0
Net Error	0	3	0	2	0	1.0
Variation	1	2	1	1	1	1.2
Bias Adjusted Error (absolute net error + variation)						2.2

In these calculations, I've assumed that the bias and variation components of error are of equal importance, so they can simply be added together. Of course, weights can be assigned to represent the relative importance of bias and variation.

By disaggregating the error calculation into a bias component and variation component, we ensure that the resulting metric picks a forecast pattern with a lower or lowest sum of bias and variation. In this example, the second forecast is now correctly identified as a better fit than the constant forecast at the mean of 4.

For completeness, we show the bias-adjusted error for the zero forecasts in the lowest frame in Table 6. The ME is -3, reflecting the tendency to underforecast by a total of 15 units and mean value of 3. Variation about this mean averages 3.6 units, and so adding the mean bias and variation yields a bias-adjusted error of 6.6 units, clearly inferior to the other two sets of forecasts.

Bias-adjusted error therefore successfully measures the error associated with intermittent-demand forecasts in a meaningful manner, thereby solving the numerator problem – the biggest problem that most practitioners didn't even realise they had!

To aggregate error metrics across products, we need a scale-free metric: to this end, the bias-adjusted error can serve as the numerator over any denominator that is not exploded by a sequence of zeros, such as the mean of the actual demand. Doing so yields a metric formally analogous to the MAD/MEAN – except that, while the MAD does not adjust for bias, the bias-adjusted variation metric builds this adjustment in.

While the bias-adjusted variation metric provides a solution to the numerator problem arising from intermittent demands, it has the added advantage of readily generalizing to situations of normal demand. This is the subject of Part 2 of this article, in the next issue of *Foresight*.

APPENDIX

How to Calculate Bias-Adjusted Mean Absolute Error

The formula for bias-adjusted mean absolute error (BAMAE) is calculated as follows, where t is a period, n the number of periods, and e the error (forecast less the actual value):

Step 1:

calculate bias (mean error):

Bias (ME) = (Σet..tn) x 1/n

Step 2:

calculate variation (mean absolute error excluding bias)

Variation (MAUE) = (Σ|(et..tn – ME)|) x 1/n

Step 3:

calculate BAMAE by adding bias expressed in absolute terms to the variation:

BAMAE = MAUE + |ME|

Steve Morlidge is the founder of Satori Partners, a source of independent help and advice for businesses wanting to change their performance management practices. He has authored numerous articles for *Foresight*, including multi-part series on guiding principles of forecasting and on evaluation of forecastability.
steve.morlidge@catchbull.com

CONCLUSION

The bias-adjusted error metric solves the numerator problem experienced when measuring the performance of intermittent-demand forecasts, a problem that has dogged academic work for many years. It is also relatively straightforward for forecasting practitioners to calculate and explain to their clients – and, as already mentioned, it properly reflects the manner in which forecast error has an impact on inventory levels. In principle, this means that it should be possible to apply it to the calculation of error where there is no intermittency of demand.

REFERENCES

Boylan, J. (2005). Intermittent and Lumpy Demand: A Forecasting Challenge, *Foresight*, Issue 1 (June 2005), 36-42.

Hanley, J., Joseph, L., Platt, R., Chung, M. & Belisle, P. 2001).Visualising the Median as the Minimum-Deviation Location, *The American Statistician*, v. 55:2, 150-152..

Hoover, J. (2006). Measuring Forecast Accuracy: Omissions in Today's Forecasting Engines and Demand-Planning Software, *Foresight*, Issue 4 (June 2006), 32-35.

Hyndman R. J. (2006). Another Look at Forecast Accuracy Metrics for Intermittent Demand, *Foresight*, Issue 4 (June 2006), 43-46.

Kolassa, S. & Schutz, W. (2010). Advantages of the MAD/MEAN Ratio Over the MAPE, *Foresight*, Issue 6 (Spring 2007), 40-43.

Kourentzes, N. (20114). Intermittent Demand Model Optimisation and Selection, *International Journal of Production Economics*, v. 156, 180-190.

Morlidge, S. (2013). How Good Is a "Good" Forecast? Forecast Errors and Their Avoidability, *Foresight*, Issue 30 (Summer 2013), 5-11.

Prestwich, S., Rossi, R., Tarim, A. & Hinch, B. (2014). Mean-Based Error measures for Intermittent Demand Forecasting, *International Journal of Production Research*, v. 52, August, 6782-6791.

Syntetos, A. & Boylan, J. (2005). The Accuracy of Intermittent Demand Forecasts, *International Journal of Forecasting*, v. 21, 303-314.

Teuntner, R. & Duncan, L. Forecasting Intermittent Demand: A Comparative Study, *Journal of the Operational Research Society*, v. 60, n. 3, 321-329.

Wallstrom, P. & Segerstedt, A. (2010). Evaluation of Foreasting Error: Measurements and Techniques for Intermittent Demand, *International Journal of Production Economics*, v. 128, 625-630.

Do Forecasting Methods Reduce Avoidable Error? Evidence from Forecasting Competitions

STEVE MORLIDGE

PREVIEW *The set of M-competitions – comparing the forecasting accuracy of two dozen common time series methods – is a landmark in our understanding of how different methods fare on a variety of data types. For example, one common procedure, the trend line extrapolation available in Excel, emerged as the least accurate of all, and probably should be considered a must to avoid. Yet, as Steve Morlidge tells us here, the implications for practitioners, especially demand forecasters, are not widely understood and quite possibly overlooked by most.*

Steve not only summarizes the key implications, he also uses a selection of data from the M3-Competition – the most recent (year 2000) and most comprehensive – to shed additional light on the bounds of forecastability: the best (and worst) forecast accuracy we can expect to achieve.

INTRODUCTION

In this past summer's issue of *Foresight*, I made a contribution to a long-running debate about forecastability (Morlidge, 2013). Specifically, I claimed that it was possible to determine what proportion of forecast error was avoidable, by reference to the naïve forecast error (that associated with a 'no change' forecast). The result, which I christened the *avoidability ratio*, was expressed in terms of the Relative Absolute Error (RAE). The RAE is calculated by dividing the sum of the absolute forecast errors (i.e., errors ignoring the sign) over a period by the sum of the absolute naïve forecast errors over the same period. (The naïve forecast errors are the sum of the period-by-period movement in the actuals and thus are, incidentally, a measure of the volatility of the data series.)

In that article, we presented evidence that the RAE would normally reach a minimum of 0.7. The implication is that forecasting methods could expect at best to reduce forecast error by about 30% below that of the naïve forecast. We showed that the 0.7 limit of forecastability was theoretically supported when our data lack trend and seasonality. Moreover, we demonstrated that while it is theoretically possible to beat an RAE of 0.7 if there are trending and other patterns to the data, few forecasts did, and that an RAE of about 0.5 seemed to represent a practical limit on what could be achieved. This may be because, while a complex signal makes it possible to deliver a lower RAE in theory, in practice the more complex the data pattern, the more difficult it is to forecast.

Just as telling, this work showed that around 30% of forecasts have RAEs above 1.0 – i.e., they are worse than the naïve forecast. This is somewhat shocking since forecasters argue that the naïve forecast should represent the upper bound of forecast error. What this appears to show is that our attempts to forecast the future often destroy, rather than add, value (Boylan, 2010).

These results got me thinking about the curious lack of rigorous testing that has been done on the efficacy of forecasting (Goodwin, 2011) – a very odd set of circumstances, indeed. Surely any decision to invest time and money into forecasting must be based on solid scientific foundations rather than blind faith, yes? Apparently – or so it seems – no.

Furthermore, having talked to forecasting practitioners around the world, I discovered that they were almost universally ignorant of the results of what little research had been conducted in this area, most noticeably the forecasting competitions organised by Spyros Makridakis. This despite the fact that this work is very well known and influential in academia, and its findings are highly pertinent to the practical job of forecasting.

This seems to me an important gap in knowledge, and one that should be addressed in a practitioner's journal like *Foresight*.

The aim of this article is to share the insights from some academic work on the efficacy of forecasting, focusing on the so-called M3-Competition (Makridakis & Hibon, 2000). In addition, I will describe the result of my own analysis of the M3 data.

HISTORY OF THE M-COMPETITIONS

Spyros Makridakis is arguably the closest the somewhat sober world of academic forecasting has to a rock star. As an ex-Olympian, he has rather greater athletic credentials than most and, despite having a thoroughly respectable academic career and being a founding member of the International Institute of Forecasters (publisher of *Foresight*), he has not been afraid of swimming against the tide of opinion or courting controversy.

A good deal of this reputation is associated with his work on measuring the performance of forecasting methods through forecasting competitions (the M-Competitions), the results of which were for many years treated like an unwelcome, nasty smell by some of his peers.

Makridakis started this line of research when he was appointed as a consultant to a Greek firm. Disappointed to discover that the business used what he regarded as primitive forecasting methods, he set out to prove that state-of-the-art techniques like Box-Jenkins analysis would do a much better job...and failed! Shaken by these findings, he repeated the test using a larger, independent sample of data and a broader range of techniques, only to end up with the same results. He then discovered that he couldn't get his work published in academic journals – not because the methodology was flawed, but because the results didn't square with the preconceptions of the editorial board. Finally, a mauling at a meeting of the Royal Statistical Society, where his competence as a forecaster was called into question, convinced him that he would have to conduct a rigorous forecasting competition in order to convince his detractors that his findings were valid.

Key Points

- Given the importance of the issue, there has been a surprising lack of rigorous testing done on the efficacy of forecasting methods, and to what extent they add value compared to naïve projections. Moreover, practitioners are almost universally ignorant of the results of what research has been conducted in this area, most noticeably the forecasting competitions organised by Spyros Makridakis, and referred to as the M-Competitions.
- The M-Competitions shed light on these key issues: Do more-complex methods outperform simple methods? How damaging is it to choose the wrong method? How does forecast accuracy deteriorate as the lead time of the forecast increases? And what is the best accuracy we can hope to achieve at various forecast horizons?
- My analysis of the monthly industry data from the M3-Competition extends the results of the M-competitions by focusing on the extent to which different methods reduce avoidable forecast error and how close they come to the limits of forecast accuracy. The results are sobering.
- One main implication is that no technique or software package will automatically 'solve' the problems that practitioners face; indeed, it could make matters worse if applied in a thoughtless manner. Avoiding methods that are patently inadequate for any given data series can make a big difference in forecasting performance.

The results of the first competition were published in 1982 in an article subsequently voted the most influential paper of its era (Makridakis and colleagues, 1982). Continued academic scepticism at the time, however, drove these collaborators to conduct further competitions – M2 (Makridakis and colleagues, 1993) and M3 (Makridakis & Hibon, 2000) – in order to address criticisms of the methods used. The results of these competitions essentially validated the conclusions of the original work, those being that

a) statistically sophisticated or complex methods do not necessarily provide more accurate forecasts than simpler ones;

b) the relative ranking of the performance of different methods varies according to the accuracy measures being used;

c) the accuracy of approaches that combine different forecasting methods outperforms, on average, the individual methods used; and

d) the accuracy of the methods used typically deteriorates as the forecast horizon lengthens.

In hindsight, it is easy to see why these results were unwelcome.

The discipline of forecasting is based on the assumption that the future can be predicted from past behaviour, provided that the signal buried in historical data can be separated from the noise and extrapolated into the future. Many academic careers have been built on the development of ever more sophisticated ways to do this, but the results of these competitions seem to suggest that much of this effort has been wasted.

It is not too difficult to discover why these sophisticated techniques might underperform. The kinds of complex economic systems that forecasters study tend to be unstable; in other words, their behaviour changes, often very suddenly, with the result that *the future is not like the past*. In addition, complex techniques can 'overfit' the data – in effect, they 'see' patterns in historic data that don't exist but which in fact are manifestations of random fluctuations that, by definition, aren't repeated in the future. As a result, the ability to produce a good fit to history – the way that forecasting models are often selected in commercially available forecasting applications – is a poor predictor of the ability to forecast into the future.

It is also easy to see why these results might be very important for practitioners, particularly those working in the supply chain.

Supply chain professionals are often faced with the challenge of attempting to forecast demand for large numbers of products with dynamic demand patterns, and are called upon to do so very frequently. In addition, they are usually focused on improving forecast performance measured in a very particular way (related to the cost of supply) over a very specific horizon (related to the replenishment lead time). As a result of the scale of the challenge facing them, they invest large sums of money in sophisticated forecasting software, usually based on the assumption that this will lead to better outcomes. In my experience, this assumption is rarely validated in a rigorous way, using out-of-sample forecasts rather than in-sample fit to test performance.

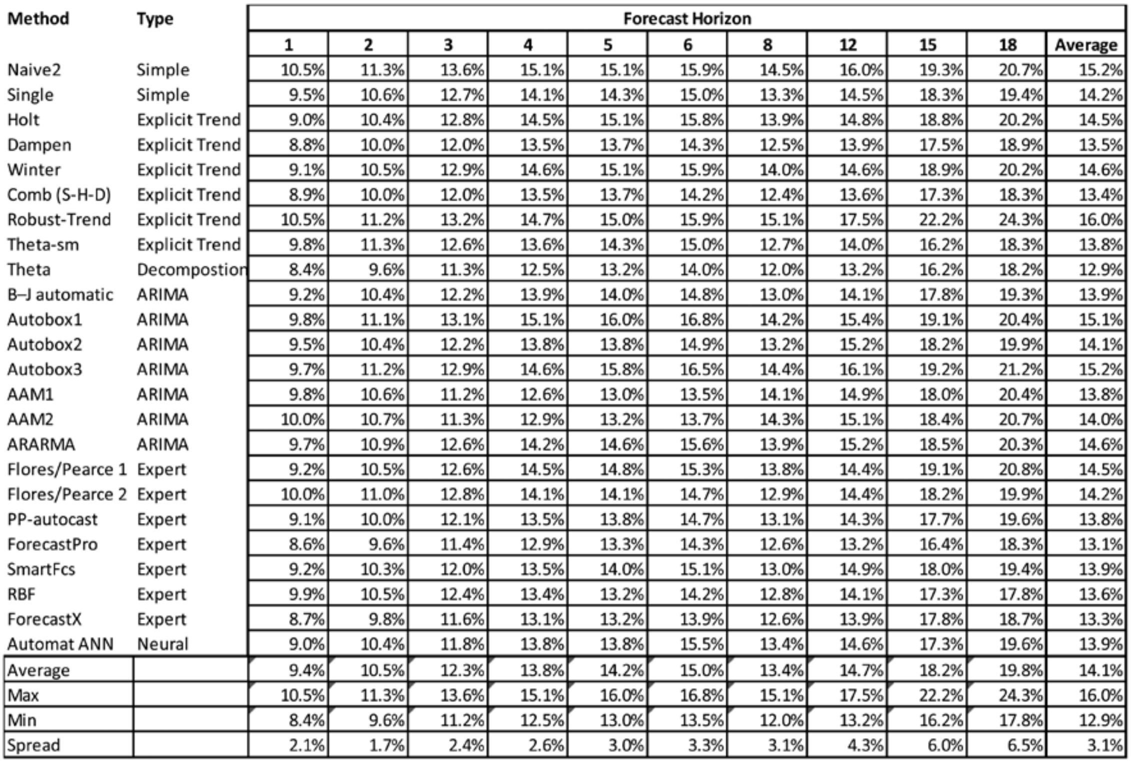
In order to tease out the implications for practitioners and to work out what they should do going forward, we need to dive into the detail, focusing on the most recent, comprehensive exercise: the M3-Competition.

THE M3-COMPETITION

The M3-Competition used 3,003 different data series – a combination of yearly, quarterly, and monthly classified as 'Micro', 'Industry', 'Macro', 'Finance', 'Demographic', and 'Other'. Each of these was forecast by experts using one of 24 different methods. These were classified (in order of relative sophistication) as either 'Simple' (e.g. Naïve 2, which is a version of the naïve method applied to seasonal data), 'Explicit Trend' models (e.g. Holt-Winters), 'Decomposition', 'ARIMA/ARARMA' (e.g. Box-Jenkins), 'Expert System' (e.g. using optimising algorithms to select the best model), and 'Neural Network'. For each series, a portion of the data was held out of the estimation process used to generate the forecasts, and then applied to assess the accuracy of these forecasts.

The M3 study produced a multitude of different analyses, but **Table 1** gives a flavour of the results. The table shows the average symmetrical MAPE (sMAPE: calculated by taking the average absolute error and dividing it by the average of the forecast and actual) over a range of horizons. As one would expect, the average error increases across the horizons, but the difference between the lowest and the highest error is not enormous; on average, the spread is three percentage points. Of the best-performing methods, some were relatively sophisticated (e.g. Forecast Pro and Theta) but others were relatively simple. The same applied to the worst-performing methods. Conclusion: sophistication is no guarantee of performance.

Table 1. Symmetrical MAPE for All Methods



Method	Type	Forecast Horizon										
		1	2	3	4	5	6	8	12	15	18	Average
Naive2	Simple	10.5%	11.3%	13.6%	15.1%	15.1%	15.9%	14.5%	16.0%	19.3%	20.7%	15.2%
Single	Simple	9.5%	10.6%	12.7%	14.1%	14.3%	15.0%	13.3%	14.5%	18.3%	19.4%	14.2%
Holt	Explicit Trend	9.0%	10.4%	12.8%	14.5%	15.1%	15.8%	13.9%	14.8%	18.8%	20.2%	14.5%
Dampen	Explicit Trend	8.8%	10.0%	12.0%	13.5%	13.7%	14.3%	12.5%	13.9%	17.5%	18.9%	13.5%
Winter	Explicit Trend	9.1%	10.5%	12.9%	14.6%	15.1%	15.9%	14.0%	14.6%	18.9%	20.2%	14.6%
Comb (S-H-D)	Explicit Trend	8.9%	10.0%	12.0%	13.5%	13.7%	14.2%	12.4%	13.6%	17.3%	18.3%	13.4%
Robust-Trend	Explicit Trend	10.5%	11.2%	13.2%	14.7%	15.0%	15.9%	15.1%	17.5%	22.2%	24.3%	16.0%
Theta-sm	Explicit Trend	9.8%	11.3%	12.6%	13.6%	14.3%	15.0%	12.7%	14.0%	16.2%	18.3%	13.8%
Theta	Decompostion	8.4%	9.6%	11.3%	12.5%	13.2%	14.0%	12.0%	13.2%	16.2%	18.2%	12.9%
B–J automatic	ARIMA	9.2%	10.4%	12.2%	13.9%	14.0%	14.8%	13.0%	14.1%	17.8%	19.3%	13.9%
Autobox1	ARIMA	9.8%	11.1%	13.1%	15.1%	16.0%	16.8%	14.2%	15.4%	19.1%	20.4%	15.1%
Autobox2	ARIMA	9.5%	10.4%	12.2%	13.8%	13.8%	14.9%	13.2%	15.2%	18.2%	19.9%	14.1%
Autobox3	ARIMA	9.7%	11.2%	12.9%	14.6%	15.8%	16.5%	14.4%	16.1%	19.2%	21.2%	15.2%
AAM1	ARIMA	9.8%	10.6%	11.2%	12.6%	13.0%	13.5%	14.1%	14.9%	18.0%	20.4%	13.8%
AAM2	ARIMA	10.0%	10.7%	11.3%	12.9%	13.2%	13.7%	14.3%	15.1%	18.4%	20.7%	14.0%
ARARMA	ARIMA	9.7%	10.9%	12.6%	14.2%	14.6%	15.6%	13.9%	15.2%	18.5%	20.3%	14.6%
Flores/Pearce 1	Expert	9.2%	10.5%	12.6%	14.5%	14.8%	15.3%	13.8%	14.4%	19.1%	20.8%	14.5%
Flores/Pearce 2	Expert	10.0%	11.0%	12.8%	14.1%	14.1%	14.7%	12.9%	14.4%	18.2%	19.9%	14.2%
PP-autocast	Expert	9.1%	10.0%	12.1%	13.5%	13.8%	14.7%	13.1%	14.3%	17.7%	19.6%	13.8%
ForecastPro	Expert	8.6%	9.6%	11.4%	12.9%	13.3%	14.3%	12.6%	13.2%	16.4%	18.3%	13.1%
SmartFcs	Expert	9.2%	10.3%	12.0%	13.5%	14.0%	15.1%	13.0%	14.9%	18.0%	19.4%	13.9%
RBF	Expert	9.9%	10.5%	12.4%	13.4%	13.2%	14.2%	12.8%	14.1%	17.3%	17.8%	13.6%
ForecastX	Expert	8.7%	9.8%	11.6%	13.1%	13.2%	13.9%	12.6%	13.9%	17.8%	18.7%	13.3%
Automat ANN	Neural	9.0%	10.4%	11.8%	13.8%	13.8%	15.5%	13.4%	14.6%	17.3%	19.6%	13.9%
Average		9.4%	10.5%	12.3%	13.8%	14.2%	15.0%	13.4%	14.7%	18.2%	19.8%	14.1%
Max		10.5%	11.3%	13.6%	15.1%	16.0%	16.8%	15.1%	17.5%	22.2%	24.3%	16.0%
Min		8.4%	9.6%	11.2%	12.5%	13.0%	13.5%	12.0%	13.2%	16.2%	17.8%	12.9%
Spread		2.1%	1.7%	2.4%	2.6%	3.0%	3.3%	3.1%	4.3%	6.0%	6.5%	3.1%

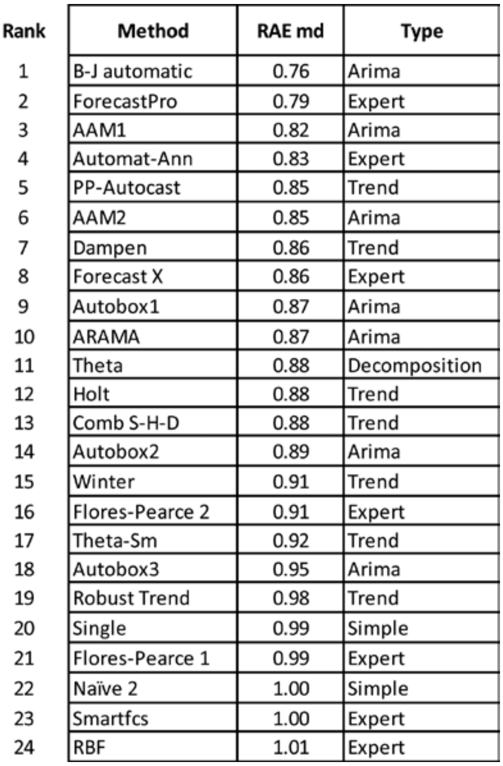
MONTHLY INDUSTRY DATA

Since my focus is on supply chain applications, I have narrowed the scope of my analysis to concentrate on the monthly 'industry' data over the short term. By this I mean lags 1 to 3, since few industries have many products with replenishment lead times longer than three months. In total, this comprises 334 data series.

We earlier made the case for using Relative Absolute Error (RAE) as the primary measure of forecast performance, since this takes 'forecastability' into account – another key consideration for supply chain practitioners.

Table 2 shows the median RAE (the median has been used to discount the large number of outliers that would otherwise distort the analysis) for all 24 methods for forecasting one month ahead. (The median figures were calculated from the raw forecast errors.) It shows that all the RAEs fall within the expected range of 0.7 and 1.0. While the spread of results is again not great, the significance of the variation in performance is. The RAE of 0.76 earned by B-J Auto, for example, lies toward the limits of what our avoidability

Table 2. The Median RAE for Lag 1



Rank	Method	RAE md	Type
1	B-J automatic	0.76	Arima
2	ForecastPro	0.79	Expert
3	AAM1	0.82	Arima
4	Automat-Ann	0.83	Expert
5	PP-Autocast	0.85	Trend
6	AAM2	0.85	Arima
7	Dampen	0.86	Trend
8	Forecast X	0.86	Expert
9	Autobox1	0.87	Arima
10	ARAMA	0.87	Arima
11	Theta	0.88	Decomposition
12	Holt	0.88	Trend
13	Comb S-H-D	0.88	Trend
14	Autobox2	0.89	Arima
15	Winter	0.91	Trend
16	Flores-Pearce 2	0.91	Expert
17	Theta-Sm	0.92	Trend
18	Autobox3	0.95	Arima
19	Robust Trend	0.98	Trend
20	Single	0.99	Simple
21	Flores-Pearce 1	0.99	Expert
22	Naïve 2	1.00	Simple
23	Smartfcs	1.00	Expert
24	RBF	1.01	Expert

Table 3. The Weighted Average RAE for Lag 1

Rank	Method	RAE wtd	Type
1	ForecastPro	0.67	Expert
2	B-J automatic	0.68	Arima
3	Dampen	0.70	Trend
4	Comb S-H-D	0.71	Trend
5	Winter	0.71	Trend
6	Forecast X	0.72	Expert
7	Holt	0.72	Trend
8	Theta	0.73	Decomposition
9	Single	0.73	Simple
10	ARAMA	0.74	Arima
11	AAM1	0.74	Arima
12	PP-Autocast	0.76	Trend
13	Autobox1	0.76	Arima
14	Autobox3	0.78	Arima
15	Naïve 2	0.79	Simple
16	Autobox2	0.79	Arima
17	Flores-Pearce 1	0.80	Expert
18	Automat-Ann	0.81	Expert
19	AAM2	0.82	Arima
20	Robust-Trend	0.86	Trend
21	Theta-Sm	0.88	Trend
22	RBF	0.88	Expert
23	Flores-Pearce 2	0.95	Expert
24	Smartfcs	0.96	Expert

Figure 1. The Distribution of the RAE

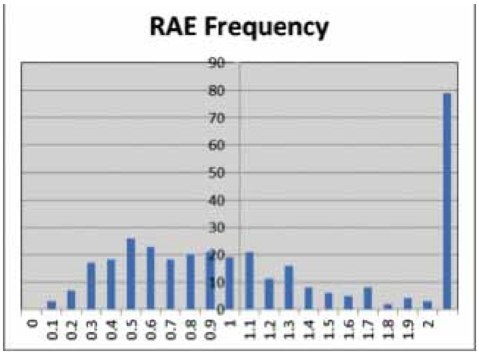
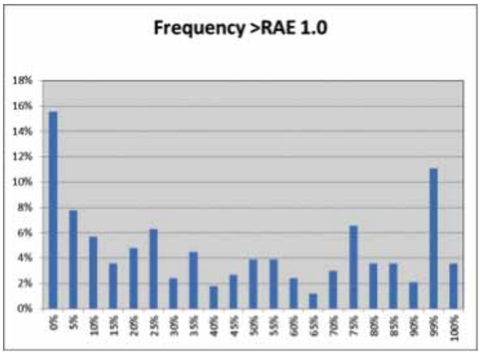


Figure 2: The Distribution of Poorly Forecast Data Series



ratio would lead us to believe is achievable, whereas the six methods with RAE of 0.98 or worse (three of which are classified as 'Expert' methods) are barely better than the simple naïve forecast.

This doesn't tell the whole picture, though. Supply chain professionals are not interested in the median RAE. What matters to them – what drives their costs – is the total amount of avoidable error. Achieving near-perfect forecast performance for a low-volume product is of less interest than slightly improving the quality of forecasting for a high-volume product. When we weight the RAE for the same data set by volume, we get a different picture, as **Table 3** shows.

The overall performance is marginally better across the board, but what is most significant is some of the changes in relative performance. In particular, nearly 50% of the methods have very good RAE of 0.75 or better. In addition, the two very simple naïve approaches are now mid-table rather than towards the bottom.

There is clearly a danger of overinterpreting these results; the data in the M3 study are not drawn from the same company, so weighting the RAE may not be meaningful (although it does illustrate a point). In addition, although the data set may be large in the academic context it is trivial in an industrial context. Also, we are analysing the results for 335 data series forecast only once for each lag; even a small business would generate a much larger sample of short-term forecast performance data than this within a few months. Nevertheless, comparing the two tables demonstrates the validity of one of the key findings from the M competitions: that the relative performance of different forecasting techniques depends upon the *metric used* to measure forecast accuracy. It is therefore very important to use metrics that are tightly aligned to the purpose of the forecast process.

All these analyses measure the average performance of different methods. What is probably more significant than the average performance of any single method is the range of performance of the methods, since it is unlikely (and inadvisable) for a business to rely solely on a single method. The results are very revealing.

Figure 1 shows the average RAE in forecasting one month ahead over all 334 monthly industry data series and across all forecasting methods.

The picture here is similar to that which we normally see when we look at the pattern of forecast performance: although the average RAE may be respectable (0.78 here), a significant proportion of data series are forecast worse, on average, than if a simple naïve forecast had been used – in this case, somewhere near 60%. But is this result because some data series are inherently unforecastable? The large number of RAE in excess of 2 is a consequence of the fact that many data series show little or no change in the period under review. They therefore have a very low naïve forecast error and very high RAE.

Figure 2 shows the proportion of times one of the forecast methods has an RAE above 1.0 in forecasting one month ahead.

This shows that a mere 16% of data series were never poorly forecast by any method and only 4% were always poorly forecast by every method. The other 80% fall somewhere in between. In other words, although some methods are better than others, and some data series are easier to forecast than others, it suggests that it is possible to forecast most series badly if you choose the wrong method. It also seems that choosing the wrong method is a pretty easy thing to do. In the case of the M3 study for Lag 1, every single method fared worse than naïve between 32% and 50% of occasions.

Again, there is a danger of overinterpreting the results – but they are consistent with the empirical work I have carried out to date, and so seem to be a real phenomenon rather than a statistical aberration. In addition, the same result holds true for Lags 2 and 3.

IMPLICATIONS FOR PRACTITIONERS

The implications here for practitioners are profound yet in a sense self-evident. What is most important for improving forecast performance is not optimising the forecasting method, but rather avoiding methods that are patently inadequate. If all data series with an RAE of greater than 1.0 were forecast using the naïve forecast (thus forcing the RAE back to 1.0), the average weighted RAE for all methods in the M3 would improve from 0.78 to 0.58, close to the limit of what is achievable in practice (0.5).

This simple analysis therefore suggests that the benefit of identifying and eliminating circumstantially poor forecasting methods could be nearly as great as the benefit from using 'sophisticated' forecasting methods on their own. Since we do not yet have the capability to identify the potential for poor forecasting before the event, the only way that any of this performance improvement can be achieved is by routinely and rigorously measuring actual forecast performance after the event (in a meaningful way), and taking remedial action as soon as it becomes clear that the level of performance is significantly suboptimal.

CONCLUSIONS

This paper discusses some of the important empirical work on the relative performance of different forecasting methods, work of which practitioners are largely ignorant, despite it being very relevant for their purposes. In particular, this research has proved beyond reasonable doubt that sophistication of method is no guarantee of performance, and that what constitutes 'performance' is highly dependent on the metrics being used and the horizons involved. The implication is that no technique or software package will automatically 'solve' the problems that practitioners face; indeed, matters could, and probably will, be made worse if forecasting methodologies are applied in a thoughtless manner.

Further analysis of the data carried out for this paper builds on these conclusions, in that it suggests that the average performance of any method is less important than the distribution of its performance – any forecasting method will destroy value (perform worse than a simple naïve forecast) a significant proportion of the time if it is used indiscriminately.

At a practical level, it suggests that users should focus less on trying to optimise their forecasting process than on detecting where their process is severely suboptimal and taking measures to redress the problem. And the only way in which this can be done is by measuring performance in a way that is relevant to their purpose. In a nutshell: to add value to real businesses, forecasting needs to be more evidence based than theory driven.

REFERENCES

Boylan, J. (2009). Toward a More Precise Definition of Forecastability, *Foresight*, Issue 13 (Fall 2009), 34-40.

Goodwin, P. (2011). High on Complexity, Low on Evidence: Are Advanced Forecasting Methods Always as Good as They Seem? *Foresight*, Issue 23 (Fall 2011), 10-13.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R. & Al, E. (1982). The Accuracy of Extrapolation (Time Series) Methods — Results of a Forecasting Competition, *Journal of Forecasting*, 1, 111-153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, D. K. & Al, E. (1993). The M2-Competition — A Real Time Judgementally Based Study, *International Journal of Forecasting*, 9:1, 5-22.

Makridakis, S. & Hibon, M. (2000). The M3 Competition: Results, Conclusions, and Implications, *International Journal of Forecasting*, 16, 451-476.

Morlidge, S. (2013). How Good Is a "Good" Forecast? Forecast Errors and Their Avoidability, *Foresight*, Issue 30 (Summer 2013), 5-11.

Steve Morlidge is author of *Foresight*'s multiple-part series on the Guiding Principles of the Forecasting Process (2011-2012). His analysis of the forecastability issue is an ongoing project. If you would like to participate or contribute data, please contact him at:

steve.morlidge@catchbull.com