CHAPTER 1

Fundamental Considerations in Business Forecasting

Challenges in business forecasting, such as increasing accuracy and reducing bias, are best met through effective management of the forecasting process. Effective management, we believe, requires an understanding of the realities, limitations, and principles fundamental to the process. When management lacks a grasp of basic concepts like *randomness*, *variation*, *uncertainty*, and *forecastability*, the organization is apt to squander time and resources on expensive and unsuccessful fixes: There are few other endeavors where so much money has been spent, with so little payback.

This chapter provides general guidance on important considerations in the practice of business forecasting. The authors deal with:

- Recognition of uncertainty and the need for probabilistic forecasts
- The essential elements of a useful forecast
- Measurement of forecastability and bounds of forecast accuracy
- Establishing appropriate benchmarks of forecast accuracy
- The importance of precisely defining *demand* when making demand forecasts
- Guidelines for improving forecast accuracy and managing the forecasting function

■ ■ ■

Although we were unable to secure rights to include it in this book, Makridakis and Taleb's "Living in a World of Low Levels of Predictability" from the *International Journal of Forecasting* is an important piece worth mentioning in any consideration of fundamental issues.

Spyros Makridakis is very well recognized as lead author of the standard forecasting text, *Forecasting: Methods and Applications*, and of the M-series forecasting competitions. Through his books, *Fooled by Randomness* and *The Black Swan*, Nassim Nicholas Taleb has drawn popular attention to the issue of unforecastability of complex systems, and made "black swan" a part of the vernacular. Their article, published in the *International Journal of Forecasting* (2009), speaks to the sometimes disastrous consequences of our *illusion of control*—believing that accurate forecasting is possible.

While referring to the (mostly unforeseen) global financial collapse of 2008 as a prime example of the serious limits of predictability, this brief and nontechnical article summarizes the empirical findings for why accurate forecasting is often not possible, and provides several practical approaches for dealing with this uncertainty. For example, you can't predict when your house is going to burn down. But you can still manage under the uncertainty by buying fire insurance.

So why are the editors of a *forecasting* book so adamant about mentioning an article telling us the world is largely unforecastable? Because Makridakis and Taleb are correct. We should not have high expectations for forecast accuracy, and we should not expend heroic efforts trying to achieve unrealistic levels of accuracy.

Instead, by accepting the reality that forecast accuracy is ultimately limited by the *nature* of what we are trying to forecast, we can instead focus on the efficiency of our forecasting processes, and seek alternative (nonforecasting) solutions to our underlying business problems. The method of forecast value added (FVA) analysis (discussed in several articles in Chapter 4) can be used to identify and eliminate forecasting process activities that do not improve the forecast (or may even be making it worse). And in many situations, large-scale automated software can now deliver forecasts about as accurate and unbiased as anyone can reasonably expect. Plus, automated software can do this at relatively low cost, without elaborate processes or significant management intervention.

For business forecasting, the objective should be:

> *To generate forecasts as accurate and unbiased as can reasonably be expected*—and to do this as efficiently as possible.

The goal is not 100% accurate forecasts—that is wildly impossible. The goal is to try to get your forecast in the ballpark, good enough to help you make *better decisions*. You can then plan and manage your organization effectively, and not squander resources doing it.

1.1 GETTING REAL ABOUT UNCERTAINTY*

Paul Goodwin

Business forecasters tend to rely on the familiar "point" forecast—a single number representing the best estimate of the result. But point forecasts provide no indication of the uncertainty in the number, and uncertainty is an important consideration in decision making. For example, a forecast of 100 ± 10 units may lead to a much different planning decision than a forecast of 100 ± 100 units.

In this opening article, Paul Goodwin explores the types of "probabilistic" forecasts, the academic research behind them, and the numerical and graphical displays afforded through prediction intervals, fan charts, and probability density charts. Providing uncertainty information, he explains, can result in better decisions; however, probabilistic forecasts may be subject to misinterpretation and may be difficult to sell to managers. There is also an unfortunate tendency we have to seriously underestimate the uncertainty we face and hence overstate our forecast accuracy.

Goodwin's article provides practical recommendations and additional sources of guidance on how to estimate and convey the uncertainty in forecasts.

* This article originally appeared in *Foresight: The International Journal of Applied Forecasting* (Spring 2014), and appears here courtesy of the International Institute of Forecasters.

Avoiding Jail

In October 2012, the scientific world was shocked when seven people (engineers, scientists, and a civil servant) were jailed in Italy following an earthquake in the city of L'Aquila in which 309 people died. They had been involved in a meeting of the National Commission for Forecasting and Preventing Major Risks following a seismic swarm in the region. At their trial, it was alleged that they had failed in their duty by not properly assessing and communicating the risk that an earthquake in the area was imminent. Their mistake had been that they had simply conveyed the most likely outcome—no earthquake—rather than a probabilistic forecast that might have alerted people to the small chance of a strong earthquake (Mazzotti, 2013).

Point versus Probabilistic Forecasts

This case dramatically highlights the problem with forecasts that are presented in the form of a single event or a single number (the latter are called *point forecasts*). They give no information on how much uncertainty is associated with the forecast. As such, they provide no guidance on what contingency plans you should make to cope with errors in the forecasts. Is the risk of an earthquake sufficient to evacuate an entire town? How much safety stock should we hold in case demand is higher than the forecast of 240 units?

But incorporating uncertainty into forecasts is not straightforward. Probabilistic forecasts need to be presented so that they are credible, understandable, and useful to decision makers—otherwise, we are wasting our time. And, as we shall see, getting reliable estimates of uncertainty in the first place poses its own challenges.

Prediction Intervals

Prediction intervals are a common way of representing uncertainty when we are forecasting variables like sales or costs. The forecast is presented as a range of values, and the probability that the range will enclose the actual outcome is also provided. For example, a 90% prediction interval for next month's demand for a product might be given as 211 to 271 units (or 241 ± 30 units). Clearly, the wider the interval, the greater uncertainty we have about the demand we will experience next month.

Fan Charts

More information about uncertainty is provided by a fan chart (see Figure 1.1). Here, the darkest band represents the 50% prediction interval, while the wider ranges show the 75% and 95% intervals.

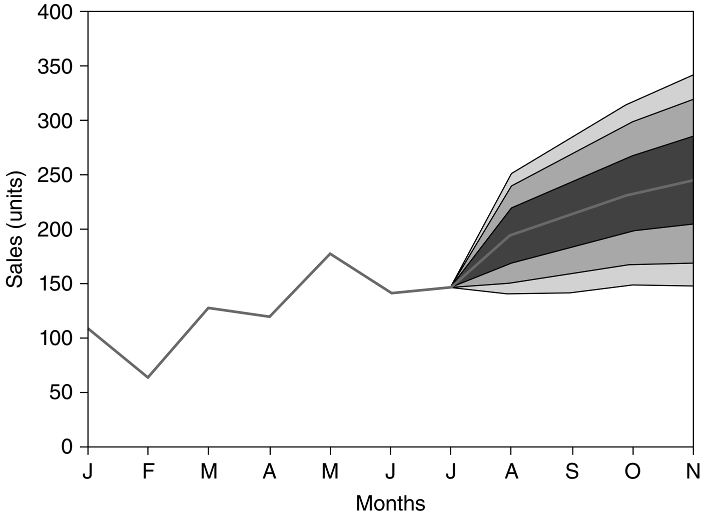

Figure 1.1 A Fan Chart

Probability Density Chart

Lastly, the forecast can be presented as an estimate of an entire probability distribution. For example, we might forecast a 10% probability of snow, a 20% probability of rain, and a 70% chance of fine weather for noon tomorrow. Estimates of probability distributions for variables like sales, costs, or inflation are usually referred to as density forecasts. Figure 1.2 provides an example. It can be seen that sales should almost certainly fall between 200 and 1,200 units, but sales around 500 units are most probable.

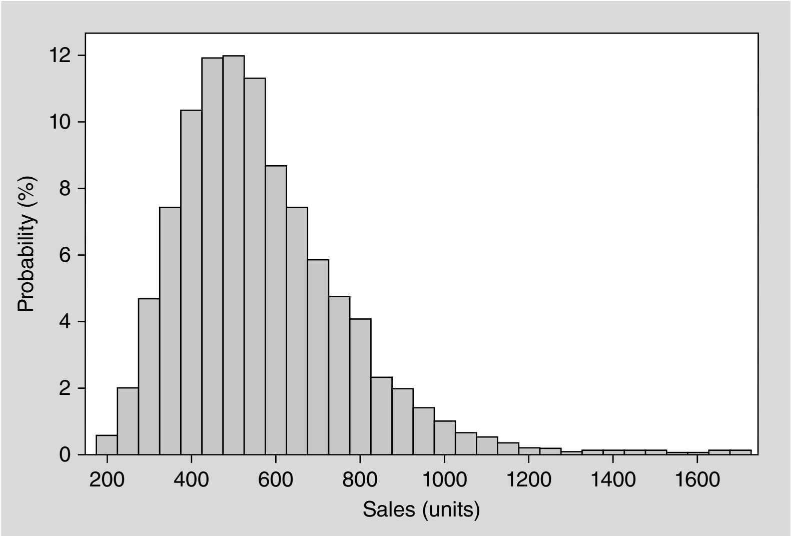

Figure 1.2 A Density Forecast

Is it Worth Communicating Uncertainty?

Does communicating uncertainty in forecasts lead to better decisions? In a recent experiment conducted by Ramos and coauthors (2013), people received forecasts of a river's level and had to make decisions on whether to open a floodgate to protect a town. Opening the gate would cause flooding of agricultural land downstream and liability for compensation payments from farmers. The decision makers received either a single-figure (point) forecast, or they were additionally given a prediction interval (e.g., 3.52 ± 0.51 meters), together with a forecast of the probability that the town would be flooded. Providing uncertainty information resulted in better decisions in that, over a series of trials, more money was lost when no uncertainty information was available.

Clear advantages of providing prediction intervals are also evident in research carried out by Savelli and Joslyn (2013). Participants provided with 80% prediction intervals for high and low temperatures in Seattle were more decisive when faced with the dilemma of whether to issue temperature warnings about freezing or very hot conditions than those provided only with point forecasts. They were also better at identifying unreliable forecasts and expected a narrower range of outcomes—so they had a more precise idea of what temperatures to expect.

Limitations of Probabilistic Forecasts

However, probabilistic forecasts are not without their limitations. In particular, they may be prone to people misinterpreting them. For example, a study carried out more than 30 years ago by Alan Murphy and coauthors (1980) found that some people interpreted a weather forecast where "the probability of precipitation today is 30%" as meaning that only 30% of the relevant region would see rain. Others thought that it meant that it would rain for 30% of the day.

Second, with interval forecasts there is often a mismatch between what you need to know for your decision and the information provided in the forecasts. For example, a 90% prediction interval for demand of 211 to 271 units does not tell you what your safety stock needs to be to achieve a 98% customer service level (Goodwin and coauthors, 2010). A density forecast would give you this information because it would present the full probability distribution—but these are not regularly available in commercial forecasting software.

Third, there can be a problem in selling probabilistic forecasts to users. An interval forecast may accurately reflect the uncertainty that is being faced, but it is likely to be spurned by decision makers if it is too wide and judged to be uninformative. For example, a 90% prediction interval for sales of 50 to 900 units will probably be regarded as useless. Worse still, it is likely to cast doubts on the competence of the forecaster who produced it.

Sometimes, the reactions of users may be more nuanced. Ning Du and coauthors (2011), in a study of earnings forecasts, found that when people recognized there was significant uncertainty in what was being predicted, interval forecasts carried more credibility than point forecasts. However, only a limited interval width was tolerated. Wider intervals that were judged to be uninformative had less credibility.

What Is the Best Way to Convey Uncertainty?

All of this indicates that we need to know more about how to convey uncertainty to forecast users. Some recent studies offer a few pointers. One of these (Kreye and coauthors, 2012) provided experts on cost estimation with graphical forecasts of the monthly price of a raw material that were given in three different forms: a line graph showing minimum, maximum, and medium estimates; a bar chart showing the same information; and a fan chart. Fan charts were found to be the most effective method for raising awareness of the uncertainty that was present.

Another study, this by David Budescu and coauthors (2012), found that the uncertainty associated with forecasts produced by the Intergovernmental Panel on Climate Change (IPCC) were best communicated to the public using both words and numerical probabilities together. For example, an event might be referred to in a report as “likely, that is having a probability of 67% to 90%” or “very unlikely, that is having a probability of 2% to 10%.” Supplying valuations of uncertainty using only words, such as “likely” or “virtually certain,” was less effective. People seeing both words and numerical probabilities were more consistent in interpreting the messages, and—importantly—their interpretations were closer to what the authors of the report intended.

When prediction intervals are provided, it appears that users trust them more (in the sense that they make smaller judgmental adjustments to them) if their bounds are expressed in everyday language like “worst-case forecast” and “best-case forecast” (Goodwin and coauthors, 2013). Trust can also be increased by supporting prediction intervals with scenarios or narratives that provide a justification for their two bounds (Önkal and coauthors, 2013).

Estimating Uncertainty

Even if we can provide meaningful probabilistic forecasts to users, we still have to estimate the level of uncertainty. The main problem with prediction intervals, fan charts, and density forecasts is that they tend to underestimate uncertainty. This is particularly true when the forecasts are based on managerial judgment. Research has repeatedly shown that people produce prediction intervals that are far too narrow, and thus outcomes occur outside the interval more often than they should according to the stated probability. For example, Itzhak Ben-David and coauthors (2013) reported recently that a large sample of senior U.S. financial executives produced 80% prediction intervals of one-year-ahead stock market returns that included the actual returns only 36.3% of the time.

Spyros Makridakis and coauthors (2009) suggest a simple remedy for this problem: Once you've estimated a prediction interval, double its range! However, if you need a full density forecast, other methods might help. For example, estimating probabilities for the different factors that will influence sales, rather than estimating probabilities for sales themselves, may make the judgment task easier. Monte Carlo simulation can then be used to combine these separate estimates to generate a density forecast for sales, as shown in the *Foresight* article, "Assessing Uncertainty in New-Product Forecasts" (Guthrie and Markland, 2010).

Many past studies have found that statistical methods also tend to produce overly narrow ranges for possible outcomes, although new algorithms are faring better. George Athanasopoulos and coauthors (2011) compared the performance of different forecasting methods on over 1,300 tourism time series, and found that both an automated algorithm embedded in a commercial software package and an automated algorithm for implementing exponential smoothing produced prediction intervals that were very well calibrated when the data were monthly or quarterly. For example, 95% prediction intervals contained the actual outcome around 95% of the time. Researchers are also working to enhance statistical methods for producing density forecasts (e.g., Machete, 2013).

Conclusions

Psychologists tell us that our brains seek certainty in the same way that we crave food and other basic rewards. Uncertainty is often experienced as anxiety, and can even be felt as a form of pain. But this doesn't mean that it's sensible or even advisable to ignore or underestimate the uncertainty we face, since there is evidence that accurate information on uncertainty can lead to better decisions. Probabilistic forecasts can provide this information, and researchers are making progress in finding the best ways to estimate and convey uncertainty so that these forecasts are more reliable, understandable, credible, and useful to decision makers.

REFERENCES

Athanasopoulos, G., R. J. Hyndman, H. Song and D. C. Wu (2011). The tourism forecasting competition. *International Journal of Forecasting* 27, 822–844.

Ben-David, I., J. R. Graham, and C. R. Harvey (2013). Managerial miscalibration, *Quarterly Journal of Economics* 128, 1547–1584.

Budescu, D. V., H-H Por, and S. B. Broomell (2012). Effective communication of uncertainty in the IPCC Reports. *Climatic Change* 113, 181–200.

Du, N., D. V. Budescu, M. K. Shelly, and T. C. Omer (2011). The appeal of vague financial forecasts. *Organizational Behavior and Human Decision Processes* 114, 179–189.

Goodwin, P., M. S. Gönül, and D. Önkal (2013). Antecedents and effects of trust in forecasting advice. *International Journal of Forecasting* 29, 354–366.

Goodwin, P., D. Önkal, and M. Thomson (2010). Do forecasts expressed as prediction intervals improve production planning decisions? *European Journal of Operational Research* 205, 195–201.

Guthrie, N., and D. Markland (2010). Assessing uncertainty in new-product forecasts. *Foresight*, Issue 16 (Winter), 32–39.

Kreye, M., Y. M. Goh, L. B. Newnes, and P. Goodwin (2012). Approaches to displaying information to assist decisions under uncertainty. *Omega* 40, 682–692.

Machete, R. L. (2013). Early warning with calibrated and sharper probabilistic forecasts. *Journal of Forecasting* 32, 452–468.

Makridakis, S., R. Hogarth, and A. Gaba (2009). *Dance with Chance*. Oxford: Oneworld Publications.

Mazzotti, M. (2013). Seismic shift. *Times Higher Education* 2121, 38–43.

Murphy, A. H., S. Lichtenstein, B. Fischhoff, and R. L. Winkler (1980). Misinterpretations of precipitation probability forecasts. *Bulletin of the American Meteorological Society* 61, 695–701.

Önkal, D., K. D. Sayım, and M. S. Gönül (2013). Scenarios as channels of forecast advice. *Technological Forecasting and Social Change* 80, 772–788.

Ramos, M. H., S. J. van Andel, and F. Pappenberger (2013). Do probabilistic forecasts lead to better decisions? *Hydrology and Earth Systems Sciences* 17, 2219–2232.

Savelli, S., and S. Joslyn (2013). The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. *Applied Cognitive Psychology* 27, 527–541.

1.2 WHAT DEMAND PLANNERS CAN LEARN FROM THE STOCK MARKET*

Charles K. Re Corr

The value of conveying uncertainty and other considerations for what makes a forecast useful to investors is the subject addressed by Charles Re Corr of Merrill Lynch. Wall Street, he observes, is generally averse to providing specific numerical forecasts, not only because accurate forecasting is difficult, but also because negative forecasts can be bad for business.

Forecasts are still necessary, however, because they are the basis for making investment decisions. But financial forecasting differs fundamentally from demand forecasting in that new

* This article originally appeared in *Journal of Business Forecasting* (Fall 2012), and appears here courtesy of Dr. Chaman Jain, editor in chief.

information can be immediately integrated into market valuations. A demand planner, on the other hand, is forced to move more slowly in reaction to new information—having to work around production and inventory schedules and human resource policies.

Acknowledging the difficulty of accurate financial forecasting, Re Corr lists seven characteristics that make a forecast useful for decision making. These are time frame (a specific period or date at which to compare forecasts to actuals), direction (up or down), magnitude (a specific number, although a distribution about that number is more valuable), probability (assigning probabilities to the distribution of possible outcomes), range (highest and lowest possible outcome), confidence (statistical or subjective), and historical forecast error. Ultimately, he concludes, a "perfect forecast" need not be 100% accurate, but should provide enough information to improve management's decisions under conditions of uncertainty.

If you think forecasting product demand is hard, try forecasting the future values of the Standard and Poor's 500 Stock Index (S&P500)! Dissimilar as they might be, there are lessons to be learned for demand forecasting from stock market forecasts.

Despite much popular evidence to the contrary, Wall Street collectively has an aversion to putting a number on the future value of the stock market. This is primarily due to three reasons:

1. It is very difficult to forecast accurately.
2. A cynic would point out that a negative forecast is not good for business.
3. The market in the past has trended upward; so is the naïve forecast, "It will go higher," calling for more of the same, which has been on average fairly correct. The statistics support the viability of a naïve forecast over the past 86 years, from 1925 to 2011. During these years, Large Cap Stocks, as represented by the S&P500, have been up over the previous year 62 times, a little better than 70%. The naïve forecast, therefore, isn't so bad even if it is woefully incomplete.

Why Forecast the Future Market

We all forecast because all decisions we make require some expectations about the future. Accurate forecasts improve our chances of making the right decision. The problem for stock market predictions is that even though the trend bias is upward, the magnitude of downward markets in any year can wipe out successes of many years.

Telling the investor that market history will repeat itself evokes the basic question, "Will it happen in my remaining lifetime?" Generally, Wall Street encourages a long-term horizon, which is termed *strategic*. Since it may take years before the forecast can be proven accurate, the more vague the forecast, the better. Ironically, any allocation decisions among different asset classes—stocks, bonds, cash, real estate, and so forth—are based on forecasts, even if not acknowledged.

Allocation percentages represent confidence levels about expected returns, risks, and correlations. Even if I weigh all classes equally (because I have no opinion or information), I am in fact expressing a forecast.

The fact is that even the gold standard certification of Wall Street analysts, the coveted designation of Chartered Financial Analyst (CFA), encourages candidates not to forecast. In one study on behavioral finance, James Moniter, a Visiting Fellow at the University of Durham and a Fellow of the Royal Society of Arts, wrote a whole section titled "The Folly of Forecasting: Ignore all Economists, Strategists, and Analysts."

Why is it so difficult? The primary reason is that the market is itself a leading economic indicator. It has been used by the government as part of its Leading Indicator Index for years, and now is reported by the Conference Board monthly. Essentially, the market predicts the economy; therefore, you are trying to "predict the predictor."

Historical analysis of the S&P500 suggests that the market moves in anticipation of economic activity, up or down, about six to nine months in advance. It also sometimes signals a change in the economic activity that does not occur. The question you might ask is this: Why is the S&P500 index such a sensitive indicator? The answer is that new information is integrated into market valuations almost immediately.

This is very different than what happens in demand forecasting. The response time for a company to react to, for example, the hint of slowing sales is not the same as that of an institutional investor receiving the same information. A company has to work around production issues, inventory levels, human resource policies, and the like before it can respond to changing markets; whereas an institutional portfolio manager, within minutes of his or her decision, only has to call the trading desk and sell/buy millions of dollars of that company's stocks or bonds, virtually instantaneously in response to new information.

What Makes Any Forecast Useful?

As difficult as it is, we have to forecast. Clients expect it, and all decisions are made based on assumptions about the future. The prize for being more successful than your competitor is worth the effort.

So, what would make up a valuable forecast? Here is a famous quote:

> *"All forecasts are wrong, some are useful."*

Although this may seem like an escape clause for forecasters, if we accept the first part of the quote as true ("All forecasts are wrong"), then we are left with the question, "What makes some of them useful?" It is the descriptive elements of a forecast provided to decision-makers that prove to be useful.

Some of these seven elements are more obvious than others, depending on the industry, but all are worth reflecting upon. They are: time frame, direction, magnitude, probability, range, confidence, and historical forecast error for similar forecasts.

Time frame: What date, or period, are you using for your ending forecast? "Soon" is not a date—you need a close of business date to compare the forecast with actual results. The time frame depends on the decision makers' cycle. Even if not requested it can be very helpful to provide a series of forecasts up to the requested end-forecast date. For example, management might want a one-year-out number, yet it would be valuable for them to see the trend in three-, six-, and nine-month forecasts. This can help them understand the forecasted trend, and possibly seasonality. This can relieve anxiety if management knows the intermediate forecast numbers may be temporarily trending in the wrong direction.

Direction: Very simply, the forecast is compared to a baseline: Will it be up or down from that on the forecasted date? As with time frame, trend and seasonality may create an intermediate point where the end forecast looks like it will be wrong because it goes down before it goes up.

Magnitude: "Up" is not an amount—you need a specific amount. Although most managers like one number because it makes it easier to make a decision, they need to be reminded that the number will be wrong. Distribution around the number is the gold standard of a forecast, which is expressed in terms of probability.

Probability: A single-number forecast is called a point forecast, and by definition is assumed to be a mid-point of the possible outcomes. Therefore, 50% of the outcomes will fall on either side of the number. You can, however, provide a higher probability of meeting or exceeding the forecast by simply reducing your forecast magnitude. Here is an example using the stock market: If you provide a point forecast that the stock market will be up 10% by the end of the next 12 months (a 50% probability), you can increase the probability of beating the target by reducing the forecast magnitude. In other words, if you believe there is a 50/50 chance that the market will rise by 10%, then there is an even higher probability it will rise by 5% (the probability depends on the distribution of possible outcomes). You may do this on your own or at the management's request to help them understand the forecast.

Range: Providing a high and a low number can be very valuable to management for decision making. Low may take you to a different direction, down versus up. This is useful because management can decide if it can live with the potentially downward outcome.

Confidence: Use statistical confidence levels if available. If necessary, provide a subjective confidence level; it is, after all, part of the forecasting job. If the confidence level is subjective, make sure to tell the decision maker. This approach can at least allow you to put a probability on a positive outcome.

Historical forecast error: In most cases the forecast you are doing is repetitive and, therefore, you have the past data. Past accuracy is informational, but past consistency is also useful. Being consistently wrong, although embarrassing, can be very valuable to decision makers, depending on the range of errors.

Some Errors Are More Forgiving than Others

It is important to recognize which components of your forecast are critical—for instance, magnitude, time frame, or direction. What about market shocks (stock market and economic conditions), natural disasters, terrorist acts, and so on? Obviously, if we could predict such things, we would be in extremely high demand. There are those who continually predict such events; we usually consider them Cassandras. If they turn out to be right, people will hold them in great esteem and ask you why you did not see that coming. But for the most part, shocks are shocks because they are not generally expected.

We can treat market shocks in a similar manner as new product launches that do not go as planned. Here we look at history to find similar events and try to draw some inferences with respect to production or inventory adjustments and how long it might take to recover.

In addition to the seven components described above, there is one more point that could also be of value to any professional forecaster. The finance industry has created a wide variety of exchange-traded funds whose movements can be helpful in forecasting a wide variety of product categories. These funds represent economic sectors and industries, and, like the S&P500, they tend to move in advance of their representative sector or industry. They are baskets of stocks that are believed to fairly represent the underlying sector such as Materials, Energy, or Healthcare sectors, which could serve as a leading indicator for industries such as North American natural resources, home construction, and pharmaceuticals.

Although attributed to many, I believe it was Yogi Berra who once said,

> *"Forecasting is difficult, particularly about the future."*

Providing complete data and continually finding ways to improve forecasts can increase your value as a professional. Ultimately, the "perfect forecast" is the one that has enough information to improve management's decisions under conditions of uncertainty.

1.3 TOWARD A MORE PRECISE DEFINITION OF FORECASTABILITY*

John Boylan

One challenge that is poorly understood, difficult to resolve, and, as a consequence, often ignored is the determination of forecastability, the potential forecasting accuracy of an item, product, or revenue flow. Forecastability is the basis of benchmarking: If we can know the best accuracy we can hope to achieve, we would have a benchmark to judge how effective our current efforts are and how much room remains for improvement. This is the subject of the next three articles.

By *forecastability,* John Boylan refers to the range of forecast errors that are achievable, on average. But, he points out, the concept of forecastability needs sharpening. Boylan shows that forecastability is not the same as stability, the degree of variation in demand over time. He argues that forecastability should be measured by a band or interval in which the lower bound is the lowest error we can hope to achieve and the upper bound is the maximal error that should occur. With such a band, we could know how far we've come (reducing error from the upper bound) and how far we can still hope to go (to reduce error to the lower bound).

Clearly, any forecasting method producing greater errors (less accurate forecasts) on average than the upper bound should be discontinued. The main difficulty, of course, lies in calculating a lower bound—how can we know the potential for forecasting accuracy?

In general, we can't pin down this lower bound. But Boylan explains that we can frequently make useful approximations of the lower bound of forecast error by relating the product to be forecast to its position in the product hierarchy, by combining forecasts from different methods, and by identifying more forecastable series.

Stability versus Forecastability

The idea of forecastability has been championed by Kenneth Kahn (2006). In fact, the term *forecastability* can be interpreted in various ways. It can relate to an assessment of the stability of a data series, as in Peter Catt's (2009) usage. It can also refer to the degree of accuracy when forecasting a time series and can indicate the precision with which we estimate an expected range for the mean absolute percentage error (MAPE) when employing a time-series method.

It's clear that the concepts of stability and forecast accuracy are related. We expect forecast accuracy to deteriorate as a series becomes less stable (more volatile). We anticipate that it is harder to estimate the expected range of any error measure as a series becomes less stable. Nonetheless, stability and forecast accuracy are distinct concepts. We should remember this in order to avoid confusions that arise from using forecastability to refer to different things.

* This article originally appeared in *Foresight: The International Journal of Applied Forecasting* (Spring 2009), and appears here courtesy of the International Institute of Forecasters.

The definition of forecastability as stability makes no reference to forecasting methods or forecast-error measures. This is a strength, as the definition then relates to the data series alone and is not restricted to any particular forecast method or error measure. But it is also a weakness, as the link between stability and forecastability isn't always apparent.

In some cases, stability and forecast accuracy align nicely. The sine wave is an example of a perfectly stable time series, with no random components. If we know the phase and amplitude of the sine series, then we can forecast the series precisely. For any sensible error measure, in this case, the forecast error will be zero.

In the Hénon map example, it is assumed that the data-generating process is known to be chaotic. If we base our assessment of its forecastability on the approximate entropy metric, we would say that the series is stable. It is only forecastable, however, in the sense of forecast accuracy if the process can be identified and the parameters estimated accurately. It is doubtful if a forecaster, presented with a short Hénon time plot, would be able to deduce the dynamical system it is based upon. If the forecaster mis-specifies the data generating process, forecast errors may be large and difficult to determine. So stability of a series does not automatically imply good forecast accuracy.

This raises the question: Is stability a necessary condition for good forecast accuracy? When a series is considered in isolation, without contextual information or accompanying series, this may be the case. A volatile series cannot be extrapolated with great accuracy. However, a volatile series may have a time-lag relationship to another series, enabling good forecast accuracy to be obtained.

Alternatively, qualitative information about the business environment may enable accurate forecasts of a volatile series using judgmental forecasting methods. So taking a perspective broader than extrapolation, we can see that stability is not a necessary condition for good forecast accuracy.

Stability is important but should be distinguished from forecastability. The term *forecastability* has been used in various ways, making the concept rather slippery. A sharper definition is required, one leaving stability out of the picture.

Defining Forecastability in Terms of Forecast Error

Tentatively, I offer this definition: "Forecastability is the smallest level of forecast error that is achievable." One series is more forecastable than another, with respect to a particular error measure, if it has a smaller achievable forecast error. To avoid technical difficulties, the word *smallest* must be interpreted sensibly, according to the forecasting error metric being used.

Three examples will show that caution is needed with this interpretation. For the mean absolute error, "smallest" simply means the "lowest." For the mean error, "smallest" means "closest to zero" (e.g., a mean error of +1 is "smaller" than a mean error of −2). For the Accumulated Forecast to Actual Ratio (Valentin, 2007), "smallest" means closest to 100 (e.g., a value of 102% is "smaller" than a value of 96%).

This definition does suffer from some problems.

The first problem is that, if we take the error measure over just one period (say, the next period), we may be lucky and forecast the value exactly, giving a forecast error of zero. Clearly, such luck is not sustainable over the long term. To overcome this difficulty, we can amend the definition of forecastability to "the lowest level of forecast error that is achievable, on average, in the long run."

This definition of forecastability is not restricted to one particular error measure but can be applied to any forecast error metric for which the word *smallest* is interpreted appropriately. Nor is this definition of forecastability restricted to a "basic time-series method" (as suggested by Kahn, 2006).

Rather, it can refer to any forecasting method. In doing so, it addresses Peter Catt's objection to the use of the coefficient of variation of a series after decomposition (removal of linear trend and seasonality). Classical decomposition, which may be considered a "basic time-series method," is just one method that can be applied to detrending and deseasonalizing the series. Perhaps, after taking into account autocorrelation based on the more complex ARIMA modeling, we may be left with a smaller coefficient of variation. My definition of forecastability overcomes this difficulty by not limiting the scope of forecasting methods that may be applied.

A second problem: The definition depends on the achievement of the smallest forecast error. It is possible that a series is difficult to forecast and will yield high forecast errors unless a particular method is identified, in which case the forecast errors are small. In cases such as these, it would be helpful to specify both a lower bound and an upper bound on forecast errors. Methods for estimating these upper bounds are discussed in the following sections.

Our definition is now broadened accordingly: "Forecastability refers to the range of forecast errors that are achievable on average, in the long run. The lower value of the range represents the lowest forecast error achievable. The upper value of the range represents an upper bound based on a benchmark forecasting method."

Upper Bound of a Forecasting Error Metric

If we could find an upper bound for forecasting error, based on a simple benchmark method, then any method producing greater errors (less accurate forecasts), on average, should be discontinued and an alternative sought. An upper bound can also be used to generate exception reports, to inform corrective actions by a forecasting analyst.

Many relative error metrics use the naïve as the benchmark method. The naïve method predicts no change from the present to the next future period. Metrics that incorporate the naïve baseline include the relative absolute error, the Theil coefficient, and the mean absolute scaled error (Hyndman and Koehler, 2006). For all of these metrics, results above 100% show that we could do better by using the naïve, the last actual observation as the forecast. Relative error measures with the naïve as the baseline are provided in most forecasting software packages.

One disadvantage of using the naïve as the upper bound is that it may set too low a bar. Often, it is obvious that better alternatives are available, especially when the data are trended or seasonal. The M1 and M3 forecasting competitions (Makridakis et al., 1982; Makridakis and Hibon, 2000) confirm that the naïve is generally inferior to other simple forecasting methods. This research evidence matches the experience of practitioners, who would be unlikely to view the naïve as a viable forecasting method.

Two alternatives may be considered. For nonseasonal data, the simple moving average or simple exponential smoothing may be used as a baseline. For trended or seasonal data, a baseline that takes trend and seasonality into account (such as classical decomposition or Winters' exponential smoothing) may be more sensible. These alternatives take the approach suggested by Kahn (2006) but use it as an upper bound, rather than as a lower bound. As Peter Catt argues, methods based on decomposition of trends and seasonal components can often be improved upon; while not appropriate as lower bounds, they can be used as upper bounds. These upper bounds should be sharper than the naïve method, meaning that analysts will be able to detect problems with current forecasting methods earlier, as they are being compared with better alternative methods.

Lower Bound of a Forecasting Error Measure

The previous section has indicated some methods for determining an upper bound on forecast accuracy. How about the lower bounds? If the data-generating process (DGP) is known, and the time series does not deviate from the DGP in the future, then it may be possible to set the lower bound exactly. This is done by determining mathematical expressions for the long-run averages (expectations) of the error measure. This approach has been adopted in studies of seasonal DGPs and is discussed later.

When we do not know the data generating process, or when the DGP is changing over time, the lower bound must be estimated. This is the situation facing the practitioner working without the luxury of well-specified, well-behaved data.

At first, the estimation of a lower bound for forecasting error may seem an impossible task. After all, there are endless forecasting methods, weighted averages (combinations) of methods, and judgmental approaches that may be used.

One approach is to estimate the lower bound of a set of methods $M_1, M_2, \ldots, M_m$. For example, M_1, M_2 may represent two methods currently used by an organization. The other methods may not be used presently but may be under consideration for implementation. But we can't be sure that we've included the ideal or optimal method in this set. So we should expect that the lower bound from our set of methods will not be the ultimate lower bound.

In Figure 1.3, I have assumed that the ultimate lower bound is unknown. We have reordered the methods so that method M_1 has the largest error, and method M_m has the smallest error. The error induced by method M_m is a measure of forecastability, when the methods are restricted to the set of methods $M_1, M_2, \ldots, M_m$.

From a more practical perspective, users of forecasting software may wish to examine the forecastability of series by using automatic model-selection procedures. Automatic forecasting is based on a set of methods built into the software, and an error measure is used to pick the best method. This approach can be applied to give an immediate lower bound, based on the software being used and an error measure of the user's choosing (not necessarily the same as the one used by the software to "pick best"). It also serves as a very useful benchmark for assessing judgmental adjustments to software-generated forecasts. If forecasts are consistently improved by the application of judgment, then the lower bound can be reduced further, giving a more accurate indication of the forecastability of the series. For example, Syntetos et al. (2009) found that a pharmaceutical company was able to improve the accuracy of its intermittent demand forecasts, based on company software, by incorporating judgmental adjustments. Thus, the lower bound had been reduced.

An alternative approach to the comparison of a set of methods is to look at combinations of those methods. For example, suppose we are considering

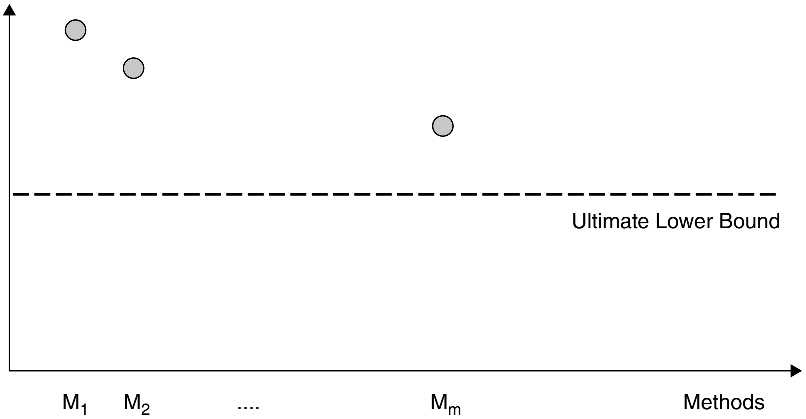

Figure 1.3 Forecast Error Lower

five methods, $M_1, M_2, \ldots, M_5$. We can also examine the simple averages of all subsets of these methods, starting with all pairs, then moving on to all triplets and so on, until we finish with the average of all five methods. The best combination can be used as our lower bound if it produces lower forecast errors than each of the methods used individually.

Armstrong (2001) argued that such combinations are particularly valuable if you use methods that differ substantially and are drawn from different sources of information. Graefe and his colleagues (Graefe et al., 2009) found that simple averaging of four components of U.S. election forecasts improved accuracy of poll share forecasts. Goodwin (2009) summarized evidence from three studies on economic forecasting, all of which showed that combining forecasts is likely to improve accuracy. The M1 and M3 competitions also showed combination methods to perform well, making them a natural choice to include in estimating the lower bound of forecast error.

In some cases, such as new-product launches, it is not possible to compare methods, or combinations of methods, based on historical data. In this situation, the best we can do is to conduct such analyses for analogous series (e.g., launch of a similar product some time ago). When more data become available, our lower (and upper) bound estimates can be refined.

Finding More Forecastable Series

One strategy for improving forecasting error has received much attention, namely working to improve statistical forecasting methods (or linear combinations of methods). A second strategy is to take advantage of judgmental forecasts or judgmental revisions to statistical forecasts. Increased attention is now being given to this important aspect of forecasting. A third strategy is to identify more forecastable series to forecast. This strategy has received less attention in the forecasting literature but has great potential to reduce forecasting errors.

Hau Lee and colleagues (Lee, 2000) studied demand at a retailer that follows a particular type of model, known as an autoregressive model of order one. In this model, the current demand is related to the demand in the previous period by a multiplicative factor, plus a random disturbance term. If the multiplicative factor is positive, then the series is said to be "positively autocorrelated."

Lee and his coauthors supposed that the retailer uses an "order-up-to" ordering policy on the manufacturer. At each review period, the retailer places an order to bring the stock up to a predefined level. In these circumstances, if demand is positively autocorrelated at the retailer, and optimal forecasting methods are used, then the orders on the manufacturer will have greater variability than the demand at the retailer. This is an example where one series (orders on the manufacturer) is inherently less forecastable than another series (demand at the retailer). It makes sense, in this case, to share information so that the wholesaler can base orders on the more forecastable demand at the retailer. There have been recent case studies showing that this strategy can reduce costs significantly in practice (Boone and Ganeshan, 2008).

Lee's model was developed by assuming that the DGP followed an autoregressive structure. In real-world applications, demand may not follow this autoregressive process, or an optimal forecasting method may not be used, or the inventory policy may not be "order-up-to." In such cases, a range of forecasting methods can be applied on retailer demand and on orders to the manufacturer. As indicated in the previous section, the errors induced by the best methods may be compared, to assess which series is more forecastable. This is a pragmatic policy, since the range of potential forecasting methods employed by many organizations is limited by considerations such as forecasting software and familiarity by forecasting analysts. Of course, we may be missing a method that would reverse our decision about which series is more forecastable. This can be addressed only by a more exhaustive method search.

Another example of finding a more forecastable series relates to seasonality. Estimation of seasonal indices is often difficult, particularly if there are few years of data and the series are noisy. In many practical situations, there is a wealth of associated data that could prove helpful. The same product may be sold at many different locations, for example. If it is reasonable to assume that the same seasonal patterns prevail at all locations, then the seasonality of the total demand may be more forecastable than the seasonality of the individual series (at the different locations). If we use a multiplicative index, then seasonal indices found at the aggregate level can be applied directly at the individual level. A similar argument applies for product families, where it is reasonable to assume that the same seasonal indices apply to all products in a family.

Leonard (2007) discusses the use of hierarchical models for seasonality, incorporating many individual time series and their aggregates. It should be noted that the aggregate series are not always more forecastable. Chen and Boylan (2007) present rules for the use of aggregate series for seasonal models, based on comparisons of expressions for the lower bound of forecast error (based on mean squared error). Suppose one series is very noisy, but its seasonality conforms to the group. Then it can "borrow strength" from the other series, and an aggregate seasonal index should be used. Suppose a second series also has seasonality that conforms to the group, but its data are very well behaved, with little noise. Then it will only "borrow weakness" from the group, and it is better to use its own individual seasonal index. In a subsequent paper, Chen and Boylan (2008) applied their rules to real data from a manufacturer of light bulbs, showing that accuracy gains can be achieved.

Conclusions

The idea of forecastability is valuable; it allows the focus of attention to shift from forecasting methods to the very series that are being forecasted. The concept of forecastability requires sharpening, however. I proposed this definition: "Forecastability refers to the range of forecast errors that are achievable on average, in the long run. The lower value of the range represents the lowest forecast error achievable. The upper value of the range represents an upper bound based on a benchmark forecasting method."

By not restricting the concept to a particular forecasting method or forecast error measure, there are two benefits. First, the concept is more general, allowing for error measures from a very broad class of error metrics. It is not restricted to basic time-series methods. Secondly, it may be applied to both theoretically generated series and to real-data series. The former may give some indication of the circumstances under which one series is more forecastable than another. The latter can be used to test such insights on real data, using forecasting methods and error measures that are relevant to the organization. This approach is well worth examining in practice, as substantial gains in forecasting accuracy may be attained.

REFERENCES

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.): *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Norwell, MA: Kluwer Academic Publishers, 413–439.

Boone, P., and R. Ganeshan (2008). The value of information sharing in the retail supply chain: Two case studies. *Foresight: International Journal of Applied Forecasting* 9, 12–17.

Catt, P. (2009). Forecastability: Some insights from physics, graphical decomposition, and information theory. *Foresight: International Journal of Applied Forecasting* 13, 24–33.

Chen, H., and J. E. Boylan (2008). Empirical evidence on individual and group seasonal indices. *International Journal of Forecasting* 24, 525–534.

Chen, H., and J. E. Boylan (2007). Use of individual and group seasonal indices in subaggregate demand forecasting. *Journal of the Operational Research Society* 58, 1660–1671.

Goodwin, P. (2009). New evidence on the value of combining forecasts. *Foresight: International Journal of Applied Forecasting* 12, 33–38.

Graefe, A., J. S., Armstrong, A. G. Cuzán, and R. J. Jones (2009). Combined forecasts of the 2008 election: The Pollyvote. *Foresight: International Journal of Applied Forecasting* 12, 41–42.

Hyndman, R. J., and A. B. Koehler (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting* 23, 679–688.

Kahn, K. B. (2006). In search of forecastability. Forecasting Summit 2006, Orlando, Florida.

Lee, H. L., S. K. So, and C. S. Tang (2000). The value of information sharing in a two-level supply chain. *Management Science* 46, 626–643.

Leonard, M. (2007). Forecasting short seasonal time series using aggregate and analogous series. *Foresight: International Journal of Applied Forecasting* 6, 16–20.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandoski, J. Newton, E. Parzen, and R. Winkler (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting* 1, 111–153.

Makridakis S., and M. Hibon (2000). The M3 Competition: Results, conclusions and implications. *International Journal of Forecasting* 16, 451–476.

Syntetos, A. A., K. Nikolopoulos, J. E. Boylan, R. Fildes, and P. Goodwin (2009). The effects of integrating management judgment into intermittent demand forecasts. *International Journal of Production Economics* 118, 72–81.

Valentin, L. (2007). Use scaled errors instead of percentage errors in forecast evaluations. *Foresight: The International Journal of Applied Forecasting* 7, 17–22.

1.4 FORECASTABLITY: A NEW METHOD FOR BENCHMARKING AND DRIVING IMPROVEMENT*

Sean Schubert

Extending this discussion of forecastablity, Sean Schubert shows how to create internal benchmarks based on product-specific attributes, such as product volume, product stability, and company and market characteristics. Schubert calls these characteristics the forecastability DNA of the product, and his approach attempts to supply a prediction of the forecast error we should expect based on the item's DNA, hence providing internal benchmarks for forecast accuracy.

As a practical consequence, the forecastability model can help identify items that are "uncharacteristic" and therefore require special attention. It also allows comparison across different businesses, and determines suitably customized targets for forecast accuracy.

Introduction: Establishing Comparability

Whenever the topic of forecasting comes up in polite company, one of the first questions asked—after the complaints about bad forecasts have died down—is, "Well, what should my forecast accuracy be?" Since few seasoned business professionals believe that a target of 100% forecast accuracy is realistic, the question then becomes, "What forecast accuracy is achievable?"

When we talk about a metric or key performance indicator (KPI), we typically ask, "How are other companies doing?" Benchmarking competitors in similar businesses is certainly a relatively simple way to see how we compare.

* This article originally appeared in *Foresight: The International Journal of Applied Forecasting* (Summer 2012), and appears here courtesy of the International Institute of Forecasters.

If our competitors are at 70% accuracy (30% error on average) while we're at 50%, we know we have some work to do. If we're at 75% and similar businesses are at 65%, then we can feel confident that we're holding our own and getting results.

The devil is in the details of how these metrics are calculated in different companies, or even in different business units within a single company. Stephan Kolassa stresses this point in his *Foresight* article (Kolassa, 2008) on the usefulness of benchmarking surveys: "In benchmarking, comparability is the key!"

To establish comparability of forecasting metrics, we need to ask, among other things:

- What is the forecast lead time?
- What time buckets are we forecasting (monthly, weekly, or daily)?
- What level in the product hierarchy are we looking at (SKU, SKU × Distribution Center, SKU × Customer × Location, etc.)?
- Is anything scrubbed out of the accuracy metric?
- How is the metric weighted?
- Are direct-import or make-to-order SKUs included?

We might also question how hard a business is to forecast. Generally, it's easier to forecast sales for a nonseasonal consumer staple like diapers than the latest short-lived fad aimed at tweens. We try to neutralize this by benchmarking competitors in similar businesses, but even if we're calculating the metric consistently and the companies are in the same sector, can we say everything else—strategy, supply chain, product makeup, customer behavior, and so on—is also equal, or even reasonably similar? Without reasonable comparability on all the relevant dimensions, we agree with Stephan Kolassa that "the quest for external forecasting benchmarks is futile" (Kolassa, p. 13).

Since scratching the surface of benchmarking creates as many new questions as it answers, where does that leave us when we think about assessing forecastability?

What Is Forecastability?

Some researchers define forecastability as the ability to improve on a simple forecast, such as the naïve method (forecast is the same as the last actual). Such a definition treats the naïve method as the benchmark and hence may be an indicator of the minimum forecast accuracy to be expected. It is uninformative, however, about the best achievable forecast accuracy.

More generally, Ken Kahn (2006) called forecastability "an assessment of a data pattern's stability." The presumption here is that the more stable (less volatile) the series is, the more accurately it can be forecast. Still, it does not tell us what forecast accuracy improvement is achievable.

Looking still more broadly into data patterns, Peter Catt (2009) defines forecastability as the complexity of the underlying data-generating process, along a continuum from deterministic (can be forecast without error) through random (cannot be forecast any better than predicting "no change"). Catt's definition, like Kahn's, helps us to determine the relative forecastability of different products, but again does not readily translate into a metric that reveals what our accuracy aspirations should be.

All these definitions illuminate the problem at hand. How can we tell if we could be doing a better job forecasting a specific item? What is the reasonably achievable degree of forecast accuracy for a given SKU, group of SKUs, customer, or business?

These are not simple questions. Whatever the method we put in play, we won't know if we can do better until we try to do better—perhaps by using a more sophisticated method. And if a new forecasting algorithm is invented tomorrow, then maybe we'll get even better forecast accuracy. John Boylan's discussion in an earlier *Foresight* feature on forecastability (Boylan, 2009) examines these challenges in more detail, and offers some general guidelines for determining the lower bound to achievable forecast error. His suggestions include relating the time series to its position in the hierarchy, combining forecasts from different methods, and identifying more forecastable series with similar data characteristics.

But our objective here is to stay focused on where there are opportunities for improvement, how large those opportunities are, and how forecastability varies by business, region, customer, and item.

Forecastability DNA

Searching for answers to the question of what degree of forecast accuracy is reasonably achievable, we can think about the types of things that might affect the forecastability of a particular product. Table 1.1 offers a list of possible candidates for consideration. Most relate to the product itself, others to broader company policy, still others to the markets in which the product is sold.

While the list in Table 1.1 is not exhaustive, it can give us some insight into why certain products are easier to forecast than others. We can think of these factors as the forecastability DNA of each item. Once we understand an item's DNA, then we'll understand what drives its forecastability, and why it's different or similar to other items in the same or different businesses.

Let's walk through an example using just one DNA factor—the variability of the SOH, which is frequently measured (see Kahn, 2006, for example) by a statistic called the coefficient of variation (CoV). The CoV is traditionally

Table 1.1 Attributes of a Product's Forecastability

Yearly volume of Sales Order History (SOH)
Length of the SOH
Variability of the SOH
Intermittency of the SOH
Promotions (frequency, magnitude, repetitiveness)
Trend and seasonality
Forecast error from a naïve model
Number of customers and concentration of sales in the largest customers
Supply-chain and inventory strategies
Lead time required for the forecasts

defined as the ratio of the standard deviation of the series to the mean, and, as such, measures the percentage degree of variation in the series around the average.

The usual argument is that the higher the CoV—that is, the more a series fluctuates—the more difficult the series is to forecast. One notable exception is that series with seasonal sales patterns could be more forecastable despite their additional variation over the seasons.

Figure 1.4 plots the mean absolute percent error (MAPE) vs. the CoV for all the SKUs in a particular region for a business. This mix of SKUs includes both seasonal and nonseasonal, steady runners, and highly intermittent items. The forecasts were "automatically" generated using statistical software.

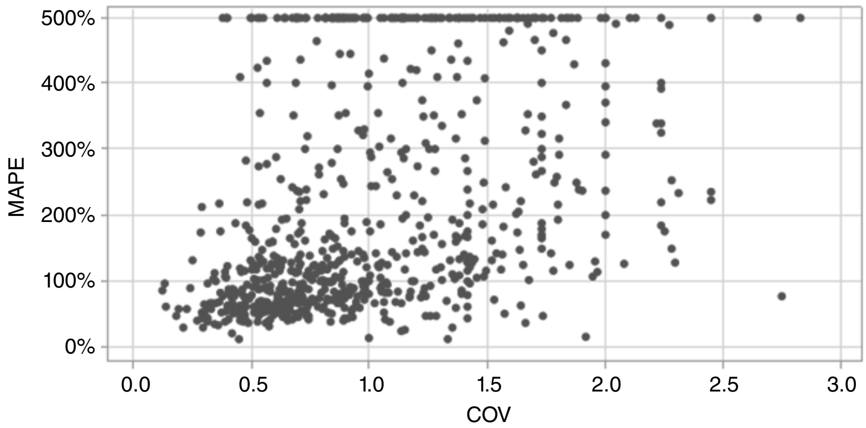

Figure 1.4 Forecast Error (MAPE) vs. Coefficient of Variation (CoV)

All SKUs for Business 2 Region 4. Forecast Error (MAPE) @ 60-day lead time. MAPEs above 500% were set to 500% to keep the graph to a manageable scale. CoV measured over recent 12 months.

In Figure 1.4, there is an apparent correlation (R^2 = 30%) that suggests that the SKUs with greater variability were forecast with greater error (less accuracy). Clearly, the CoV doesn't tell the whole story, but it can help us identify items that are "uncharacteristic" and therefore require special attention. For one thing, we see numerous items that have low CoVs (less than 0.5, for example), but still suffer high errors.

It's also logical that higher volume SKUs are easier to forecast than lower volume SKUs. Figure 1.5 shows such a correspondence (R^2 = 38%). In the same manner, we could walk through the rest of the genes in the forecastability DNA (Length of SOH, Intermittency of Sales, etc.) and select the factors that best predict forecast accuracy. However, we should use the more powerful multivariate approach that combines all that information in the DNA into a single model.

Building a Model of Forecastability

Form of the Model

Following advice from Einstein and William of Ockham, we should construct the simplest model that helps us understand forecastability, but no simpler. Once we've completed the analysis, we will get a relationship of the general form:

$$\text{Forecast Accuracy Metric} = \beta_0 + \beta_1 * \text{DNAFactor1} + \beta_2 * \text{DNAFactor2} + \beta_3 * \text{DNAFactor3} + \ldots$$

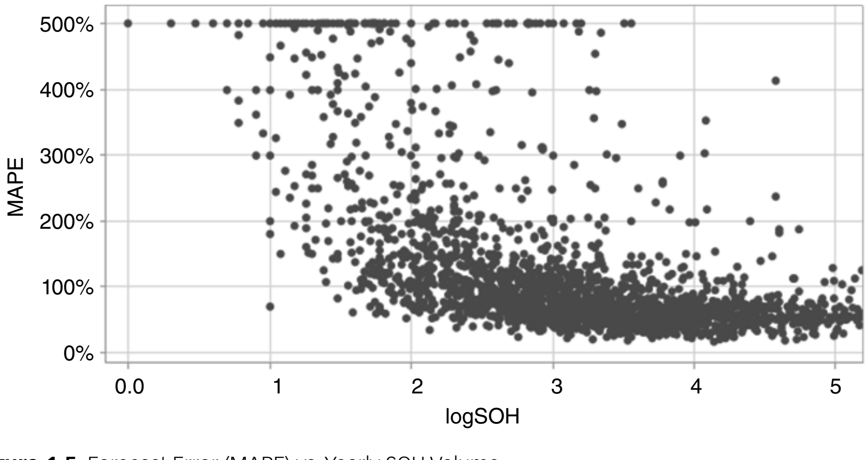

Figure 1.5 Forecast Error (MAPE) vs. Yearly SOH Volume
All SKUs for Business 3 Region 3. Forecast Error (MAPE) @ 60-day lead time. MAPEs above 500% were set to 500% to keep the graph to a reasonable scale. Yearly SOH Volume measured over recent 12 months.

With this model customized for our own unique businesses, we can proceed to benchmark ourselves against ourselves. Internal benchmarking in this way automatically adjusts the benchmark based on the specifics of each item, which addresses the comparability question as well as the common reaction to benchmarking comparisons, i.e., "My business is harder to forecast because of *x*, *y*, and *z*." By including the effects of each gene in the forecastability DNA, we address and compensate for each of those *x*s, *y*s, and *z*s, thereby developing a relevant benchmark for that business. Again, there is the caveat that while no metric is perfect, a metric is useful if it helps us improve the performance of our forecasting process.

For example, if one particular business introduces a large number of new products every year, we can compensate for that in our benchmark by including length of sales history as part of the DNA, which means we are truly comparing "like to like" across our many items, regions, customers, and businesses.

Pulling the Data Together

As presented above, an item's forecastability DNA can include quite a few genes, which implies that we need a reasonable amount of data to estimate the forecastability model. How much data is enough? A statistician would say that "more is better"—as long as it's all relevant and representative. It's best if we can get the full history, as well as additional information on customers and promotions at the SKU level; but at minimum we'll need the last 12 months of actual sales and forecasts at the item level (at key lead times of interest) with information on business, brand, product family, item, and all other details of interest. Remember, we're pulling this data together not to generate forecasts from the history but to judge and learn more about what drives the effectiveness of our forecasting process.

Internal Benchmarking

Once estimated with the available data (more is better!), the forecastability model will supply a prediction for the forecast error to be expected for each item based on its particular DNA. This prediction is our internal benchmark, which can help sound an alarm for cases where the forecasts for an item are performing significantly worse than their benchmark. In those cases we would ask questions like: "Are you using statistical forecasting?" "Are you using best practices to select and manage your forecasting models?" "Are you making inappropriate judgmental adjustments to your forecasts?" And so on.

This approach to internal benchmarking also avoids the assumption that sophisticated methods are automatically good performers: that just because a business is using advanced forecasting techniques (e.g., ARIMAX, neural nets, etc.), there is little room for improvement. Of course, we would generally expect regions or businesses using more advanced tools to beat the benchmark when compared to forecasts generated from Microsoft Excel or other "primitive" methods, but we will let the analysis tell the tale of what works.

While pulling the data together to build our internal benchmarking model, we may find there are some data missing. Even if that's the case, we may still be able to create a model by using robust modeling techniques like neural nets, ensemble methods like decision trees, or other data-mining techniques, or, in the case of a regression model, by substituting the average value of any predictor where values are missing.

Another common recommendation in the modeling literature (Gelman, 2006) is to center and normalize each of the predictor variables (*x*s) to help us create more easily interpretable model coefficients. This means for each potential predictor like CoV, we subtract the mean from all the values and divide by the standard deviation of all the values. Once we've done this for all the potentially predictive factors in the forecastability DNA, we will be ready to build our multivariate model.

Defining the Forecast-Accuracy Metric

A key decision in building the model is the definition of the forecast-accuracy metric. Should the metric measure forecast *accuracy* or forecast *error*? (See the note by David Hawitt in the Summer 2010 issue of *Foresight* for more discussion.) In addition, we must decide whether the metric should assess the absolute forecast errors (expressed in volume units) or percentage forecast errors. A problem with percentages is that, while they allow easy comparisons between SKUs, they can disguise the true impact of forecast error on the supply chain. After all, the business is much less concerned about a 50% forecast error for a SKU that sells 100 units than for a SKU that sells 1,000,000 units (assuming similar pricing, of course).

We have chosen here to express the metric using absolute forecast errors, which allows us to more directly see the impact and cost of forecast error on the supply chain. Moreover, once we have the model output in terms of absolute errors, we can readily convert the results to percentage errors (MAPEs), forecast accuracy percentage (100% minus the MAPE), or any other version of the metric (symmetric MAPE and so on) that we could choose to present our forecasting performance to the business.

Transforming the Variables

As a final "trick" in building the model, we used a log transformation for the variables that are expressed in units, have a large range of values, or a highly skewed distribution. The log transformation helps ensure that the model satisfies certain key regression assumptions, reduces the influence of the very largest values, guarantees positive values for absolute error, and facilitates interpretations of the statistical results (Gelman and Hill, 2006; sections 3.4 and 4.4 explain the role of the log transformation in more detail).

Now our forecastability model will have the general form:

$$\text{log(absolute Forecast error)} = \beta_0 + \beta_1 * \text{log(DNAFactor1)} + \beta_2 * \text{log(DNAFactor2)} + \beta_3 * \text{DNAFactor3} + \ldots$$

Certain predictors, such as the CoV, are not transformed into logs since they are ratios or percentages to begin with.

The Forecastability Model in Action

In Table 1.2, we show a portion of the regression results obtained by fitting a forecastability model to 12 months of data on all items (70,000+) in a company's global product hierarchy.

Because the list of potential drivers is long, it is likely there will be some overlap (collinearity) between and among some factors. So the final forecastability model may use only a subset of these factors.

In our illustration, the dominant factors were the Yearly SOH Volume, Naïve Error, and CoV. The #Customers and Length of History, while statistically significant, don't appear to be major drivers of forecast error in this business. Most dominant is the SOH variable: A change of one standard deviation in the SOH variable is predicted to bring about almost a 0.9 standard deviation change in the forecast error metric, which is more than three times the effect of the Naïve Error variable and more than six times the effect of the CoV.

Table 1.2 The DNA Factors and Their Impact on Forecast Error

Global Regression Results	
DNA Factor	**Regression Coefficient**
Log(SOH)	0.8869
Log(Naïve error)	0.2908
CoV	0.1328
Log(#Customers)	0.0094
Length of SOH	0.0080
R-Square	91.5%

Notes:
All factors displayed were statistically significant.
All predictors have been centered and standardized (subtract the overall mean and divide by the standard deviation). This simplifies the determination of relative factor importance.
Only selected factors from the full model have been disclosed, since the detailed forecastability model is not transportable from one business to the next. Building a forecastability model for your particular business will provide more insight than reproducing the model shown.

Interpreting the SOH Coefficient

Previously, we showed a graph for one particular business, supporting the idea that as the yearly SOH volume increases, forecast error tends to decrease as a percentage of volume. At first glance, that negative relationship appears to be contradicted by the positive coefficient of 0.8867 for the SOH variable. The apparent contradiction is explained by the difference between the use of absolute error metric in the model and the percentage error (MAPE) metric in Figure 1.5. While a larger SOH volume leads to a larger absolute error, errors as a percentage of SOH volume decline. The calculation is shown in Table 1.3. Note the last two columns showing that as the SOH variable increases, the absolute error increases but the MAPE declines.

So long as the coefficient on the log-transformed SOH factor is less than one, higher SOH volumes lead to decreasing percentage errors and thus greater forecast-accuracy percentages.

The coefficient of variation (CoV) also appears as a significant factor (as shown in Table 1.2), but clearly not a dominant one. So while our results do affirm the pattern in Figure 1.4 that higher CoV items are associated with higher forecast errors and hence diminished forecast accuracy, the CoV cannot stand on its own as an indicator of forecastability.

Using the Model for Benchmarking

Figure 1.6 compares the (log of the) actual absolute forecast errors with those predicted by the forecastability model, for all SKUs and all businesses across all regions. The actuals on the *y*-axis are from current forecasting methods in use. These forecasting methods spanned the gamut, from advanced ERP planning modules to standalone forecasting software, Excel-based forecasting techniques to manual judgment forecasts. As an aside, we could also have included forecasting software as a DNA factor in the model to learn about how our various tools (or lack thereof) improve forecast accuracy (error), but we will leave that as a potential topic for future discussion.

Table 1.3 Building Intuition about MAPE and the SOH Coefficient

SOH	log(SOH)	log(AbsErr) = 0.8867*log SOH	AbsErr = 10^(logAbsErr)	Forecast Error (MAPE%)
10	1	0.8867	7.7	77.0%
100	2	1.7734	59.3	59.3%
1,000	3	2.6601	457.2	45.7%
10,000	4	3.5468	3,522.1	35.2%
100,000	5	4.4335	27,133.1	27.1%
1,000,000	6	5.3202	209,025.9	20.9%
10,000,000	7	6.2069	1,610,274.8	16.1%

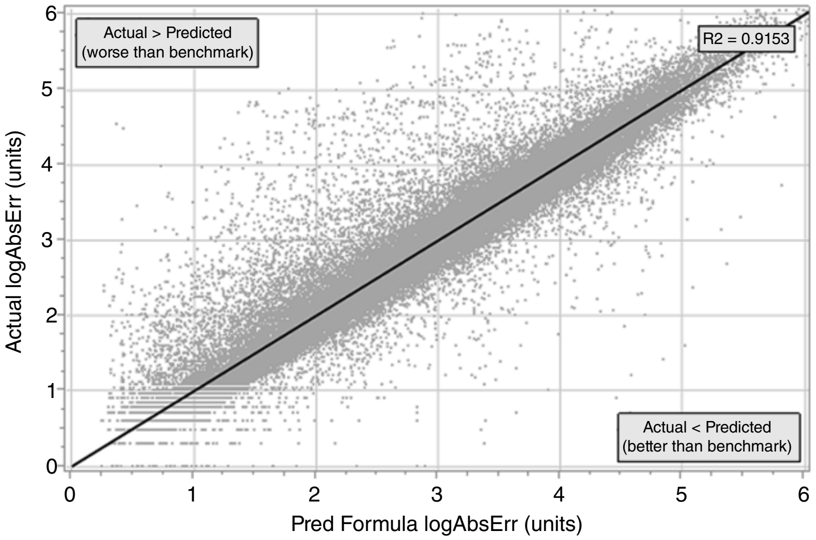

Figure 1.6 Actual vs. Predicted Benchmark Forecasting Errors (log units); All SKUs, All Businesses, All Regions
Note: The actual and predicted absolute forecast error (log units) are both measured over a 12-month period and reflect in-sample regression results.

The *x*-axis represents the predicted forecast error based on the item's forecastability DNA, that is, how difficult the item is to forecast. Higher predicted forecast error means "harder to forecast." Hence the *x*-axis represents the benchmark forecast error. Points in the area above the 45-degree line represent items for which the actual forecast error is larger than what would be predicted based on the benchmark. Points in this region suggest there may be an opportunity for improvement. We would still need to dig into the details to see if these worse-than-benchmark results are due to improper use of statistical models, non-value-added adjustments to the forecast, timing issues in the metric, inability to forecast promotions or new products, or something else awry with the nuts and bolts of our forecasting process.

Points in the area below the 45-degree line identify items for which the actual forecast error is less than the benchmark forecast error. Such results indicate that our forecast errors are lower than should be expected based on the model of forecastability DNA. While comforting, this does not necessarily imply that there is no further room for improvement. In these cases (actual less than predicted), we could raise the target by using the benchmark accuracy from one of the better-performing businesses as a goal. The mechanics of this process will be left as a potential topic for a future article.

Single-Item Benchmarking

Let's apply the forecastability model to one specific item. This item, used in the construction industry, was introduced 18 months ago: Recent forecasting errors totaled 18,554 units (absolute errors at a 60-day lead time summed over the last 12 months). Converted into percentages, the MAPE is 91.7%, implying a forecast accuracy of only 8.3%.

For this item, the DNA inputs were:

- Length of Sales Order History = 18 months
- Yearly Volume of SOH (units: last 12 months) = 20,224
- CoV (over last 12 months) = 0.58
- Naïve Forecast Error (units: last 12 months) = 11,384
- Number of Customers = 8
- Top 2 Customers = 58.4% of Sales
- And other DNA factors . . .

Loading these DNA factors into the forecastability model will give us a benchmark forecast accuracy for this SKU. We can then compare the current results (MAPE of 91.7%) to the benchmark MAPE to see if there is potential opportunity for improvement.

The forecastability model yields:

- Forecast error (units) = 9,763 (MAPE = 48.3%)
- Forecast accuracy benchmark = 100 − 48.3 = 51.7%

So our forecasts for this SKU are performing markedly worse than the benchmark. Indeed, they are less accurate than the naïve-model forecasts. The next step is to dig into the history of this SKU (Figure 1.7) to diagnose what we may have missed.

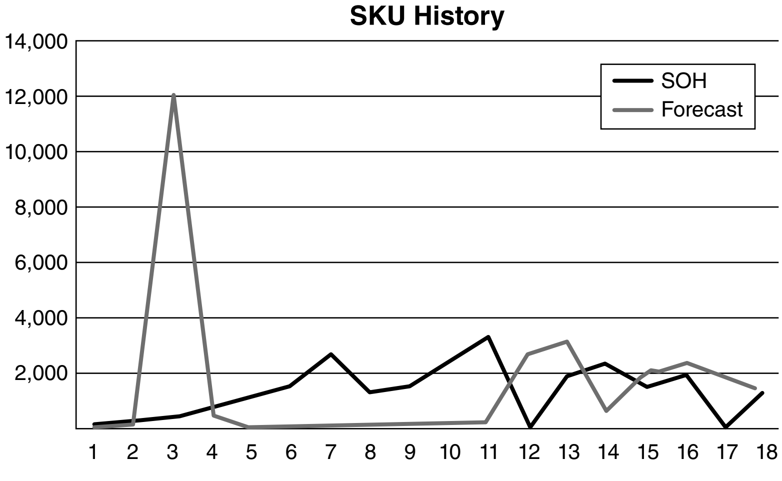

Figure 1.7 Review of Single-Item History

Note: The forecastability model uses the most recent 12 months of forecasts and actuals for modeling purposes. The full history is shown give an overview and context for discussion of forecasting over the SKU's life cycle.

We can see that the SOH has had a relatively low amount of month-to-month variation over the last 12 months. With a CoV of 0.58, it certainly should be much more forecastable than its recent forecast MAPE of 91.7%. The difficulty in forecasting appears to have been primarily driven by a lack of updating the forecast during months 5–11. Eventually the SKU sold through its initial inventory and ran out of stock during the 11th month, which affected sales and forecast error in the 12th month. The lack of sales in the 17th month requires a deeper investigation into causes; was it a brief change in the market or some type of data error in the ERP system? These are just a few of the ideas inspired by digging into the details, which in turn suggests a few areas for potential process improvement:

1. Implement tracking signals to ensure that any sustained differences between forecast and actuals are highlighted and quickly corrected.
2. Deploy a new-product-launch dashboard to ensure closer tracking of new products during initial 12 months of sales.
3. Create a system alert that highlights any SKUs where current forecast is greater than 0, recent SOH is greater than 0, but last month's SOH equals 0.

It is possible that current methods of continuous improvement for forecasting would have identified this particular SKU by its very high forecast MAPE—but an advantage to benchmarking using the forecastability-DNA approach is that it will also identify SKUs that are currently performing well using conventional metrics (e.g., ~40% forecast error), but could be performing even better (30% or less).

The power of the multivariate forecastability model also extends to understanding differences in forecastability between businesses while allowing us to generate customized forecast-accuracy targets for each business. We'll walk through an example now.

My Business Is Harder to Forecast

Knowing the key factors from the forecastability DNA for our overall business (Yearly Total SOH, Naïve Forecast Error, CoV, etc.) allows us to compare businesses objectively and quantitatively. We can start by looking at the average values of these factors for specific businesses at the SKU level, shown in Table 1.4.

Comparing Business 4 to Business 10 in the two listed regions, we see that an average SKU in Business 4 has higher volume, sells to fewer customers (Customer Count), and has similar length of history and variability (measured by CoV) as the typical SKU in Business 10. We would also note that Business 4 has half as many SKUs that need to be forecasted.

Table 1.4 Average Levels of the DNA Factors at the Business Region Level

Forecastability DNA Factor	Business 4, Region 3	Business 10, Region 2
SOH	68,689.25	13,638.65
NaiveFcstErr	30,395.17	7,893.98
CustomerCount	1.49	2.71
Length of History	10.43	10.53
CoV	0.96	0.94
SKU Count	349	861

Over the last 12 months, the accuracy achieved by these businesses (measured at the region level) is:

Business Unit	Forecast Accuracy (100%-MAPE)
Business 4, Region 3	34.5%
Business 10, Region 2	65.3%

Note that the forecastability model outputs the benchmark forecast error in units that we convert to forecast accuracy to allow easier comparisons between businesses.

The forecastability model crunches the DNA and gives the following as the benchmark forecast accuracy for each business:

Business Unit	Forecast Accuracy Benchmark
Business 4, Region 3	65.0%
Business 10, Region 2	60.5%

The forecastability DNA model also helps us understand objectively and in detail why we think Business 4 could markedly improve its forecast accuracy from its current level of 34.5%. The reasoning is:

- Business 4 has higher-volume items on average, which suggests it may be more forecastable overall than Business 10.
- Converting the average naïve forecast error per SKU of 30,395.17 for Business 4 into a naïve forecast accuracy of 55.7% also suggests the potential for improved accuracy in Business 4.
- By crunching all the forecastability DNA data, we see that the benchmark for Business 4 is calculated to be 65.0%, considerably better than recent forecast-accuracy results, which suggests that improvement is indicated for that business.

Compare how we objectively determined this target to the "standard" approach to target-setting. Following the traditional method, you'd likely hear, "Our current metric is 34.6%; our forecast accuracy goal next year will be 40.0%."

If we told them that a more appropriate goal is 65.0%, they might say, "My business is harder to forecast because of a, b, and c." Normally there isn't much we could offer in response, but with the forecastability DNA data in hand we can say, "Your CoV is similar to Business 10, your naïve accuracy is already 55%, and factoring in your number of customers, yearly sales volume (SOH), length of history, and the other genes in the forecastability DNA, a forecast accuracy better than 60% should be achievable." If the business is still skeptical, we could do further analysis to show how the items in their business are similar to many other items in other businesses that are already achieving higher forecast accuracies.

Conclusions

Forecastability and forecast-accuracy benchmarks are perennial topics of discussion in the forecasting world. Benchmarking across companies is a popular approach, but can be challenging because we don't usually know the details of how others calculate their metrics (lead time, level of aggregation, and what they scrub out, to name only a few of the unknowns), not to mention how difficult their business is to forecast. After all, even if we used the exact same metrics, few would argue that all businesses are exactly equal when it comes to forecasting difficulty.

The factors that vary both between and within businesses that can make forecasting more or less difficult can be considered part of the "forecastability DNA," which can be quantified and modeled in relation to forecastability and forecast accuracy (or error). Bearing in mind the dictum that "All models are wrong, but some are useful" from forecasting and statistics guru George Box, we can use the forecastability model to help us understand what makes our forecasting process tick, which can help us set realistic forecast-accuracy (or error) targets customized for the specifics of each item and area in the business, while pointing us in the direction of areas for potential improvement. Of course, the approach will not unequivocally state what the cause of the poor forecasting performance is, but it does support a management-by-exception approach to focusing where the opportunity for improvement is the greatest.

REFERENCES

Boylan, J. (2009). Toward a more precise definition of forecastability. *Foresight: International Journal of Applied Forecasting* 13 (Spring), 34–40.

Catt, P. (2009). Forecastability: Insights from physics, graphical decomposition, and information theory. *Foresight: International Journal of Applied Forecasting* 13 (Spring), 24–33.

Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Hawitt, D. (2010). Should you report forecast error or forecast accuracy? *Foresight: International Journal of Applied Forecasting* 18 (Summer), 46.

Kahn, K. (2006). In search of forecastability. Presentation at the Forecasting Summit. Orlando, FL, February.

Kolassa, S. (2008). Can we obtain valid benchmarks from published surveys of forecast accuracy? *Foresight: International Journal of Applied Forecasting* 11 (Fall), 6–14.

1.5 FORECAST ERRORS AND THEIR AVOIDABILITY*

Steve Morlidge

Several other recent articles have broken new ground in search of useful benchmarks of forecastability. New metrics have been proposed (and old ones criticized), new analytical approaches have been offered, and many new perspectives have emerged. Still, the search continues for what we might call an *industry standard*: a protocol for assessing forecastability.

Key concepts emerged from the earlier articles 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 (stability) over time and their forecastability.
- Limitations of the coefficient of variation in measuring forecastability and an alternative in a metric of entropy.

In continuing pursuit of this industry standard for assessing forecastability, Steve Morlidge addresses the question, "How good is a 'good' forecast?" Morlidge summarizes the innovations from earlier articles and the goes on to propose a simple, logical, and supportable metric to serve as the forecastability benchmark. This is a new perspective on forecastability and is a promising basis for further work on the subject.

The simplicity of the metric he creates should be very appealing to business forecasters, given that it offers a convenient way to compare forecast accuracy results across products. 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 (because there will always be an element of randomness in our data)—a significant step toward being able to make objective statements about the lowest errors that are achievable.

* This article originally appeared in *Foresight: The International Journal of Applied Forecasting* (Summer 2013), and appears courtesy of the International Institute of Forecasters.

His approach is based on the ratio of forecast errors to the errors from a naïve model, where one forecasts no change from the present to the future. While not a new idea, he shows 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.

(For additional forecastability-related literature, see the Boylan, Schubert, and Morlidge reference sections.)

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 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 recognizes 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 nonexperts.
- 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 realizing 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 on 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 optimize 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 naive 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), a "perfect" algorithm would never achieve a ratio lower than 0.7 ($\sqrt{0.5}$).*

EXAMPLE

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.

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 characterized 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 1.8.

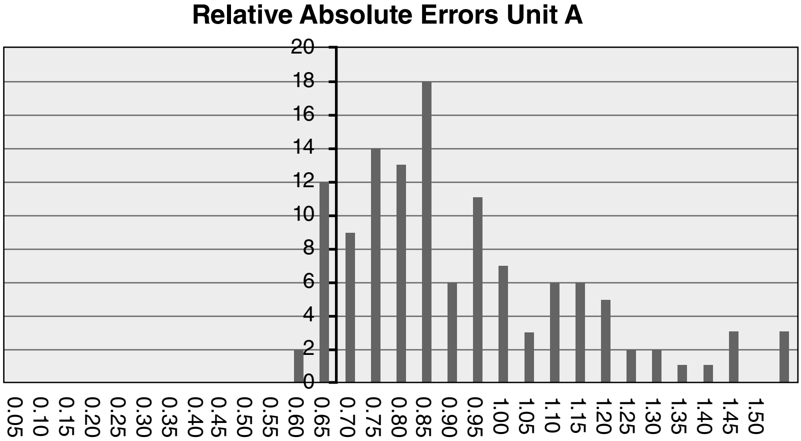

Figure 1.8 The Unavoidability Ratio (Absolute Errors Relative to Those of a Naïve Forecast) for Unit A

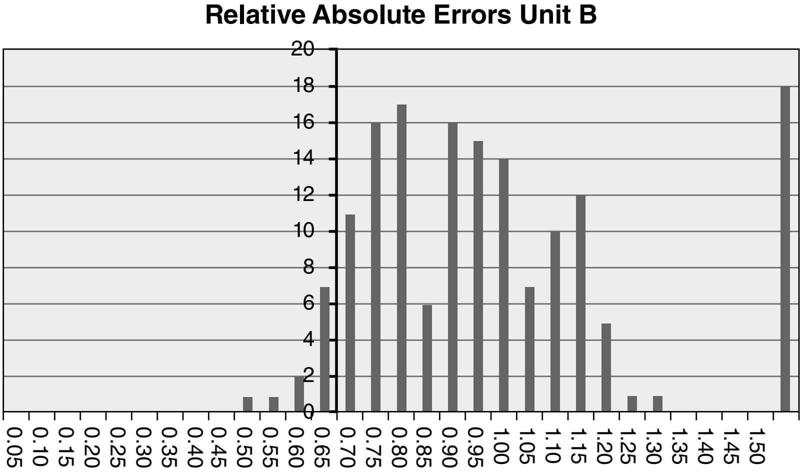

Figure 1.9 The Unavoidability Ratio (Absolute Errors Relative to Those of a Naïve Forecast) for Unit B

The second example (Unit B, Figure 1.9) 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 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 realized 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.

REFERENCES

Armstrong, S., and F. Collopy (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: International Journal of Applied Forecasting* 13 (Spring).

Catt, P. (2009). Forecastability: Insights from physics, graphical decomposition, and information theory. *Foresight: International Journal of Applied Forecasting* 13 (Spring).

Fildes, R., and Goodwin, P. (2007). Good and bad judgement in forecasting: Lessons from four companies. *Foresight: International Journal of Applied Forecasting* 8 (Fall).

Gilliland, M. (2010). *The Business Forecasting Deal*. Hoboken, NJ: John Wiley & Sons.

Gilliland, M. (2013). FVA: A reality check on forecasting practices. *Foresight: International Journal of Applied Forecasting* 29 (Spring), 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: International Journal of Applied Forecasting* 15 (Fall).

Hoover, J. (2009). How to track forecast accuracy to guide forecast improvement. *Foresight: International Journal of Applied Forecasting* 14 (Summer).

Hoover, J. (2006). Measuring forecast accuracy: Omissions in today's forecasting engines and demand planning software. *Foresight: International Journal of Applied Forecasting* 4 (June).

Hyndman, R. (2006). Another look at forecast accuracy metrics for intermittent demand. *Foresight: International Journal of Applied Forecasting* 4 (Summer).

Pearson, R. (2010). An expanded prediction realization diagram for assessing errors. *Foresight: International Journal of Applied Forecasting*, 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: International Journal of Applied Forecasting* 26 (Summer 2012).

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

1.6 THE PERILS OF BENCHMARKING*

Michael Gilliland

Organizations often seek benchmarks to judge the success of their forecasts. Reliable benchmarks allow a company or agency to see if it has improved on industry standards and to assess whether investment of additional resources in forecasting is money well spent. But can existing benchmark surveys be trusted? Do they provide useful performance standards? The next two articles consider these issues.

For benchmarking, comparability is the key to usefulness. But Michael Gilliland shows that problems can occur when the data are inconsistent, inaccurate, and unreliable—or simply inappropriate. He offers three questions for evaluating a benchmark:

1. What is the source of the benchmark data, and is it trustworthy?
2. Is the measurement consistent across all respondents?
3. Is the measure appropriate?

Appropriateness is perhaps the most important consideration, as forecasting benchmarks fail to take into consideration the underlying forecastability of each respondent's data. Gilliland also warns of the danger of blindly copying the practices of "best-in-class" companies. Their exceptional forecast accuracy may be due less to admirable practices and more to having the easiest setting for forecasting demand.

Danger, Danger

Operational performance benchmarks are available from many sources, including professional organizations, journals, and consulting and benchmarking services. Appropriately constructed benchmarks provide insight into comparative performance and can be used to guide study of the practices of companies that head the benchmark lists. But published benchmarks should not be accepted blindly because there are a number of potential perils in the interpretation of benchmark data. Problems can occur when the data are

[*] This article originally appeared in *APICS e-NEWS* (Vol. 5, No. 23, December 6, 2005), and appears here courtesy of Jennifer Proctor, editor in chief of *APICS* magazine.

inaccurate, inconsistent, and unreliable—or simply inappropriate. Here are key questions to consider.

1. What is the source of the benchmark data, and is it trustworthy?
Is the benchmark based on rigorous audits of company data or based on unaudited responses to survey questionnaires? In an audit, the competence and integrity of the auditor must be trusted. But in a survey, the trust is placed in the knowledge and motivation of all the respondents. How many people really know the answers to the questions when they are filling out the survey?

2. Is the measurement consistent across respondents?
Survey-based benchmarks are particularly troublesome when the metric is complex or ambiguous. In the forecasting realm, a simple question such as, "What is your forecast error?" requires much further specification: What is the exact error formula to use; the organizational level at which the error is measured (stock keeping unit, warehouse, customer, region, total company); the time bucket (week, month, quarter); and the lag time? Respondents may be using entirely different methods to track their errors. Even formulas as similar sounding as mean absolute percent error (MAPE), symmetric MAPE, and weighted MAPE can give dramatically different results when applied to the same data.

3. Is the measure appropriate?
One of the purposes of benchmarking is to identify top performing companies so their practices can be emulated by others. But when it comes to forecasting performance, is it really fair to compare forecast error across companies when their demand patterns may not be equally forecastable? Even within an industry, such as apparel, one company may sell long lifecycle basic items with stable demand, while another sells only "fashion" items that change every season. It would be unrealistic to expect the fashion-item forecasters to perform as well as the basic-item forecasters.

Consider this worst-case scenario:

Company ABC appears at the top of a forecasting performance benchmark list for its industry. Consultants and academics swoop down on ABC to interview management, study the forecasting process, and publish guidelines for others wishing to follow ABC's "best practices." But just because ABC has the most accurate forecasts, does it mean its forecasting process is the best or even admirable?

What if ABC had it very easy to forecast demand? Further, what if ABC's elaborate forecasting process actually made the forecast accuracy *worse* than it would have been by using a simple method such as a random walk or moving average? These are certainly not the kinds of practices that other organizations should be emulating!

In this case, the benchmark metric (forecast accuracy) was not *by itself* appropriate. Just looking at forecast accuracy did not take into consideration the underlying difficulty (or ease) of ABC's forecasting problem. It did not compare ABC's results to the results it *would have achieved* by doing nothing and just using a simple method.

An alternative metric to benchmarking is comparing the results a company achieves to the results it would have achieved by using a different method or even by doing nothing. A generalization of this forecasting approach is to conduct forecast value added (FVA) analysis. FVA is defined as the change in a forecasting performance metric (such as MAPE or bias) that can be attributed to a particular step or participant in the forecasting process. FVA helps identify process activities that are adding value by making the forecast better and also helps identify those activities that are making the forecast worse. FVA analysis is consistent with lean—helping to streamline and improve a process by identifying (and eliminating) process waste. The benefit of the FVA approach is that it can help a company get better results with less effort.

Conclusion: Beware of judging operational performance based purely on industry benchmarks. Ask the questions outlined above to assess the validity of benchmark metrics. Do not copy the so-called best practices of others without verifying that these practices are indeed adding value and the reason for improved operational performance.

1.7 CAN WE OBTAIN VALID BENCHMARKS FROM PUBLISHED SURVEYS OF FORECAST ACCURACY?*

Stephan Kolassa

Stephan Kolassa dives deeper into benchmarking surveys and argues that it is difficult if not impossible to achieve comparability through external benchmarks.

Kolassa describes the many problems that plague benchmark surveys and advises companies to redirect their search from external to internal benchmarks. Internal benchmarks provide a better representation of the processes and targets the company has in place.

Benchmarks can be trusted only if the underlying process to be benchmarked is assessed in similar circumstances. But published surveys of forecast accuracy are not suitable as benchmarks because of incomparability in product, process, time frame, granularity, and key performance indicators. A better alternative for forecast improvement is a qualitative, process-oriented target. By focusing on process improvement, forecast accuracy and the use an organization makes of the forecasts will eventually be improved.

* This article originally appeared in *Foresight: The International Journal of Applied Forecasting* (Fall 2008), and appears here courtesy of the International Institute of Forecasters.

Introduction

Sales forecasters are frequently asked what a "good" forecast is; that is, what accuracy should be expected from the forecasting method or process?

This question is important for deciding how to allocate resources to the firm's forecasting function or forecast-improvement projects. If forecast accuracy is already as good as it can reasonably be expected to be, spending additional resources would be wasteful. Thus, the company can benefit from true benchmarks of forecasting accuracy.

By true benchmarks, I mean reliable data on the forecast accuracy that can be achieved by applying best practices in forecasting algorithms and processes. Unfortunately, published reports on forecasting accuracy are rare, and those that exist suffer from shortcomings that sharply limit their validity in providing forecast-accuracy benchmarks. Consequently, I believe it is a mistake to use benchmark surveys.

Published Surveys of Forecast Accuracy

The McCarthy Survey

(McCarthy et al., 2006) studied the evolution of sales forecasting practices by conducting surveys of forecasting professionals in 1984, 1995, and 2006. Their results (see Table 1.5) provide some evidence on forecast accuracy both longitudinally and at various levels of granularity, from SKU-by-location to industry level. The forecast horizons shown are (a) up to 3 months, (b) 4–24 months, and (c) greater than 24 months. The number of survey responses is denoted by n. All percentage figures are mean absolute percentage errors (MAPEs).

One of the study's general conclusions is that the accuracy of short-term forecasts generally deteriorated over time, as shown by the weighted-average MAPEs in the bottom row. Considering the ongoing and vigorous research on forecasting, as well as vastly improved computing power since 1984, this finding is surprising. The McCarthy team conjectured that the deterioration could be due to decreasing familiarity with complex forecasting methods (as they found via interviews), product proliferation, and changes in the metrics used to measure forecast accuracy over the past 20 years.

Indeed, the survey results do suffer from problems of noncomparability. For one, the numbers of respondents in 1995 and especially in 2006 were much lower than those in 1984. In addition, I presume that the participants in 2006 differed from those in 1984 and 1995, so that lower forecast quality could simply reflect differences in respondents' companies or industries. For example, the meaning of "SKU-by-location" may have been interpreted differently by respondents in different companies and industries. Similarly, "Product Line" and "Corporate" forecasts may mean different things to different respondents.

Table 1.5 MAPEs for Monthly Sales Forecast in 1984, 1995, and 2006 Surveys

Forecast Level	Horizon								
	≤ 3 Months			4 to 24 Months			> 24 Months		
	1984	1995	2006	1984	1995	2006	1984	1995	2006
Industry	8% n = 61	10% n = 1	15% n = 1	11% n = 61	12% n = 16	16% n = 10	15% n = 50	13% n = 36	7% n = 3
Corporate	7% n = 81	28% n = 2	29% n = 5	11% n = 89	14% n = 64	16% n = 31	18% n = 61	12% n = 42	11% n = 8
Product Line	11% n = 92	10% n = 4	12% n = 6	16% n = 95	14% n = 83	21% n = 34	20% n = 60	12% n = 25	21% n = 5
SKU	16% n = 96	18% n = 14	21% n = 5	21% n = 88	21% n = 89	36% n = 36	26% n = 54	14% n = 10	21% n = 3
SKU by Location		24% n = 17	34% n = 7		25% n = 58	40% n =22		13% n = 5	
Weighted Average	15%	16%	24%						

So while the McCarthy survey provides some perspective on forecast accuracy at different times and levels, the usefulness of the figures as benchmarks is limited.

The IBF Surveys

The Institute of Business Forecasting regularly surveys participants at its conferences. The most recent survey results are reported in Jain and Malehorn (2006) and summarized in Table 1.6. Shown are MAPEs for forecast horizons of 1, 2, 3, and 12 months in different industries, together with the numbers of respondents. Jain (2007) reports on a similar survey taken at a 2007 IBF conference. The results are given in Table 1.7.

Tables 1.6 and 1.7 show large differences in forecasting accuracy among industries. For instance, the retail sector shows much lower errors than the more volatile computer/technology sector, especially for longer horizons. In general, the results show that forecast accuracy improves as sales are aggregated: Forecasts are better on an aggregate level than on a category level and better on a category level than for SKUs. And, while we should expect forecast accuracy to worsen as the horizon lengthens, the findings here are not always supportive. For example, at the Category and Aggregate levels in Consumer Products (Table 1.6), the 1-year-ahead MAPEs are lower than those at shorter horizons.

Table 1.6 MAPEs for Monthly Sales Forecast

	Horizon											
	1 Month			2 Months			1 Quarter			1 Year		
Level	SKU	Category	Aggregate	SKU	Category	Aggregate	SKU	Category	Aggregate	SKU	Category	Aggregate
Automotive	25% n = 3	5% n = 1	36% n = 1	31% n = 3	33% n = 2	25% n = 2	42% n = 1			46% n = 1		10% n = 1
Computer/ Technology	19% n = 4	14% n = 4	12% n = 7	33% n = 2	11% n = 2	18% n = 4	30% n = 3	16% n = 4	25% n = 6	17% n = 2	30% n = 1	31% n = 4
Consumer Products	27% n = 35	20% n = 23	15% n = 21	29% n = 20	22% n = 14	15% n = 10	33% n = 11	23% n = 7	14% n = 6	48% n = 4	19% n = 4	8% n = 3
Food/ Beverages	26% n = 16	15% n = 10	18% n = 11	28% n = 10	22% n = 4	36% n = 5	26% n = 8	21% n = 3	40% n = 4	19% n = 4	14% n = 2	48% n = 3
Healthcare	25% n = 7	15% n = 6	9% n = 6	27% n = 5	19% n = 5	17% n = 5	41% n = 5	24% n = 5	25% n = 5	30% n = 2	20% n = 2	15% n = 2
Industrial Products	22% n = 4	15% n = 7	7% n = 8	16% n = 2	14% n = 5	8% n = 6	17% n = 3	15% n = 6	10% n = 7	40% n = 2	21% n = 5	15% n = 6
Pharma	26% n = 5	20% n = 4	23% n = 4	30% n = 3	35% n = 2	33% n = 2	31% n = 4	25% n = 4	25% n = 3	34% n = 4	35% n = 4	28% n = 3
Retail	24% n = 7	18% n = 4	7% n = 4	17% n = 5	17% n = 6	8% n = 4	24% n = 4	10% n = 3	9% n = 4	23% n = 4	6% n = 2	6% n = 3
Telco				30% n = 1	10% n = 1	30§ n = 1	40% n = 1	15% n = 1	35% n = 1			
Others	28% n = 13	21% n = 9	17% n = 16	23% n = 7	20% n = 5	11% n = 10	25% n = 6	15% n = 5	14% n = 9	15% n = 4	18% n = 4	12% n = 8
Overall	26% n = 94	18% n = 68	13% n = 80	27% n = 58	20% n = 46	15% n = 51	30% n = 46	19% n = 37	17% n = 45	29% n = 27	21% n = 24	16% n = 33

Table 1.7 MAPEs for Monthly Sales Forecast

	Horizon											
	1 Month			2 Months			1 Quarter			1 Year		
Level	SKU	Category	Aggregate	SKU	Category	Aggregate	SKU	Category	Aggregate	SKU	Category	Aggregate
Consumer Products	29%	19%	16%	31%	20%	16%	35%	23%	22%	35%	28%	21%
Food and Beverages	27%	24%	24%	22%	12%	11%	23%	14%	15%	29%	18%	18%
Industrial Products	19%	17%	16%	28%	24%	18%	29%	22%	18%	36%	30%	17%

Unfortunately, the validity of these results is again problematic. The sample sizes were very small in many categories (Table 1.6), reflecting a low response rate by the attendees. Jain (2007) does not even indicate the number of responses behind the results in Table 1.7. In addition, these tables are based on surveys done at IBF conferences—which, after all, are attended by companies that are sensitive enough to the strategic value of forecasting to attend conferences on forecasting! Thus the MAPEs may not reflect average performance, but instead may represent lower errors at better-performing companies. Finally, while the forecast errors are shown separately for different industries—and one clearly sees large differences across industries—the industry categories are broadly defined and encompass a range of types of companies and products.

The M-Competitions

Since 1979, Spyros Makridakis and Michèle Hibon have been coordinating periodic forecasting competitions, the so-called M-Competitions. Three major competitions have been organized so far, with forecasting experts analyzing 1001 time series in the M1-Competition, 29 in the M2-Competition, and 3003 in the M3-Competition.

I will restrict the analysis here to the M2-Competition (Makridakis et al., 1993), which featured 23 series of company sales data. It attempted to model closely the actual forecasting process used in firms: Forecasters could include causal factors and judgmentally adjust statistical forecasts, and they were encouraged to contact the participating companies and obtain additional information that might influence sales. Table 1.8 shows the resulting MAPEs for monthly forecasts across different horizons, both for the average of 17 forecasting methods and for the "best" method (which I define here as the method that gave the best results, on average, across horizons up to 15 months ahead).

Table 1.8 MAPEs for Monthly Sales Forecast

Company	Industry	Number of Series	Forecast	1 Month	2 Months	1 Quarter	1 Year
Honeywell	Residential construction	6	Average Best (Naive method including seasonality)	N/A N/A	16.6% 5.1%	15.9% 6.7%	19.3% 13.5%
Squibb	Pharma	7	Average Best (Smoothing with dampened trend)	N/A N/A	9.1% 7.3%	10.6% 7.2%	28.1% 23.0%
Car company	Automotive	6	Average Best (Smoothing with dampened trend)	10.1% 8.0%	10.7% 9.5%	14.6% 14.6%	13.9% 14.2%
Aussedat-Rey	Paper	4	Average Best (Combination of smoothing methods)	3.7% 2.8%	5.6% 5.9%	6.8% 6.7%	5.2% 3.8%

The table reveals that forecast accuracy varied considerably across the four companies on a 1-year horizon, the best method yielding a MAPE of 23% for the pharma data and 3.8% for the paper data. The authors attributed the variations to different seasonalities and noise levels in the data, with pharma sales fluctuating much more strongly than paper sales. Unsurprisingly, forecast accuracy generally deteriorated as forecast horizons increased. Finally, quite simple methods—a naïve forecast, exponential smoothing with a dampened trend, or a combination of smoothing methods—beat more complex methods, including human forecasters using market information and judgmental adjustments. In particular, the Honeywell dataset showed that a simple, seasonally adjusted naïve method could be more accurate than other methods that were more complex.

However, even the results of the M2-Competition are problematic candidates for forecasting benchmarks. These companies represent a very small sample of industries, and the sample contains only one company per industry. In addition, very few time series per company were considered; for example, the only Honeywell series included were channel sales of a safety device and fan control. The latter makes it problematic even to extrapolate, from the MAPEs on the series chosen, the accuracy achievable for other Honeywell products.

Another problem is that very different series are being averaged. For instance, the six series for the car manufacturer include not only sales of three individual models (without specification of whether sales were national or international), but also total company sales and the total of the entire car industry. Conceivably, a method may forecast well for the entire automobile industry but break down when forecasting sales of a single model—a situation where life cycles need to be taken into account, although they may be less important on the aggregate level.

Finally, even though forecasting experts were encouraged to contact the companies for additional explanation and data, some experts consciously decided not to. They doubted that a sufficient understanding of the companies' markets could be formed within a short period (". . . it was hard to know what questions we should ask. . . ."). Subsequently, they acknowledged that their forecast was "not comparable with the likely accuracy of a judgmental forecast prepared within a business organization" (Chatfield et al., 1993).

Makridakis and colleagues never intended the results of the M-Competitions to be used as benchmarks against which forecasting performance of companies should be measured. Instead, the M-Competitions aimed at comparing different forecasting algorithms on standardized datasets. Their failure to provide benchmarks does not mean the results are uninformative to practicing forecasters. On the contrary, they guide practitioners to consider relatively simple methods when seeking to improve their methodologies.

What Is a Benchmark?

The concept of benchmarking is widely applied in business fields, from process benchmarking and financial benchmarking to IT performance benchmarking of new hardware. Common to any such endeavor is that measures of performance in similar and comparable fields are collected and analyzed in order to gain an understanding of what the best possible performance is.

In benchmarking, comparability is the key! Benchmarks can only be trusted if the underlying process to be benchmarked is assessed in similar circumstances. For instance, benchmarking profitability across "firms in general" fails the criterion of comparability; biotech and utility companies have widely different "normal" profitabilities, and using the best-in-class profitability of a biotech firm as a target for a utility is unrealistic.

Benchmarking is closely related to the search for best practices. Ideally, one would identify a performance benchmark and then investigate what factors enable achievement of the benchmark (Camp, 1989). For instance, an optimal sales forecast may be a result of very different factors: a good process for data collection, a sophisticated forecasting algorithm, or simply a clever choice of aggregating SKUs across stores and/or warehouses.

Any approach that leads to consistently superior forecasting performance would be a candidate for best practices. As forecasters, our search for benchmarks is really only part of our search for best practices. We try to optimize our forecasts and need to understand which part of our processes must be improved to reach this goal.

Problems with Forecast Accuracy Surveys

Can published figures on sales forecasting accuracy serve as benchmarks? My analysis indicates that the survey results suffer from multiple sources of incomparability in the data on which they are based. These include differences in industry and product, in spatial and temporal granularity, in forecast horizon, in metric, in the forecast process, and in the business model.

Product Differences. Going across industries or even across companies, we have to forecast sales of wildly dissimilar products. Sales of canned soup and lawn mowers behave very differently; their forecasting challenges will be different, too. A manufacturer of canned soup may be faced with minor seasonality as well as sales that are driven by promotional activities whose timing is under the manufacturer's control. Lawn mower sales, however, will be highly seasonal, depending crucially on the weather in early summer. Thus, it's reasonable to expect lawn mower sales to be more difficult to forecast than canned soup sales and to expect that even "good" forecasts for lawn mowers will have higher errors than "good" forecasts for canned soup.

The comparability problem arises when both canned soup and lawn mowers are grouped together as consumer products or products sold by the retail industry. This is nicely illustrated by the differences between the company datasets in the M2-Competition (Table 1.8). In addition, as I noted above, separate products of a single company may vary in forecastability. A fast-moving staple may be easily forecastable, while a slow-moving, premium article may exhibit intermittency—and consequently be harder to forecast.

Forecasts, moreover, are not only calculated for products, but also for services and/or prices. For manpower planning, a business needs accurate forecasts for various kinds of services, from selecting products for a retailer's distribution center to producing software. And in industries where price fluctuation is strong, forecasting prices can be as important as forecasting quantities. Problems of comparability may apply to price forecasts as well as to quantity forecasts. Although most published surveys have focused on quantities of non-service products, we can clearly see that benchmarking forecasts of services and prices face similar challenges.

> **Spatial Granularity.** Published accuracy figures do not precisely specify the level of "spatial" granularity. When it comes to SKU-by-location forecasts, are we talking about a forecast for a single retail store, a regional distribution center (DC), or a national DC? Forecasting at all three locations may be important to the retailer. Forecasts at the national DC level will usually be of most interest to the manufacturer, as this is the demand from the retailer he normally faces—unless, of course, the manufacturer engages in direct store delivery (DSD), in which case he will certainly be interested in store-level sales and, it logically follows, store-level forecasts.

Aggregating sales from the retail stores serviced by a regional or national DC will usually result in more stable sales patterns. Consequently, forecasting at the retail store will usually be much harder than for the national DC. A given forecast error may be fine for a store forecast but unacceptably large for a DC forecast. Similarly, it will be easier to forecast car sales of General Motors in a mature and stable market, compared to car sales by a smaller company like Rolls-Royce, which builds limited runs of luxury cars for sale to aficionados.

> **Temporal Granularity.** The time dimension of the forecasts reported in the surveys is often vague. Are the forecasts calculated for monthly, weekly, daily, or even intradaily sales? Forecasts for single days are important for retailers who need to replenish shelves on a daily basis, while weekly forecasts may be enough for supplying regional DCs. Manufacturers may only need to consider monthly orders from retailers' national DCs, but once again, in the case of DSD, they will need to forecast on a weekly or even daily level.

Just as aggregation of store sales to DC sales makes forecasting easier at the DC than in the store, it is usually easier to forecast monthly than weekly sales, easier to forecast weekly sales than daily sales, easier to forecast daily sales than intradaily sales. A given accuracy figure may be very good for a daily forecast but very bad for a monthly one.

Longer-term forecasting is harder than shorter-term, simply because the target time period is farther into the future. And long-range forecasts may differ in temporal granularity from short-range forecasts: Often, a retailer forecasts in daily (or even intradaily) buckets for the immediate next few weeks, on a monthly basis for forecasts 2–12 months ahead, and in quarterly buckets for the long term. These forecasts correspond, respectively, to operational forecasts for store ordering and shelf replenishment, to tactical forecasts for distribution center orders, and to strategic forecasts for contract negotiations with the supplier.

This example clearly illustrates that forecasts with different horizons may have different purposes and different users and be calculated based on different processes and algorithms. It's important to note that errors on different time horizons may have different costs: An underforecast for store replenishment will lead to an out-of-stock of limited duration, but an underforecast in long-range planning may lead a retailer to delist an item that might have brought in an attractive margin.

> **Key Performance Indicators (KPIs).** The published surveys employ the MAPE—or a close variation thereof—as the "standard" metric for forecast accuracy. In fact, there is little consensus on the "best" metric for sales forecast accuracy. While the MAPE is certainly the most common measure used in sales forecasting, it does have serious shortcomings: asymmetry, for one, and error inflation if sales are low. These shortcomings have been documented in earlier *Foresight* articles by Kolassa and Schütz (2007), Valentin (2007), and Pearson (2007), who proposed alternative forecast-accuracy metrics. Catt (2007) and Boylan (2007) go further, encouraging the use of cost-of-forecast-error (CFE) metrics in place of forecast-accuracy metrics.

Because of the proliferation of forecast-accuracy metrics, you can't be certain if survey respondents have actually correctly calculated the metric reported.

Then there's the asymmetry problem. Overforecasts (leading to excess inventory) and underforecasts (lost sales) of the same degree may have very different cost implications, depending on the industry and the product. Excess inventory may cost more than lost sales (as with short-life products like fresh produce, or high-tech items that quickly become obsolete), or it can be the other way around (e.g., for canned goods or raw materials). The MAPE and its variants, which treat an overforecast of 10% the same as an underforecast of 10%, may not adequately address the real business problem. KPIs that explicitly address over- and underforecasts may be more meaningful to forecast users.

Forecast Horizon. Most studies report the forecast horizon considered; I wish all of them did. Many different forecast horizons may be of interest for the user, from 1-day-ahead forecasts for the retailer to restock his shelves, to 18-months-ahead (and more) forecasts for the consumer-product manufacturer who needs to plan his future capacity and may need to enter into long-term contractual obligations.

Forecast Processes. Forecasting accuracy is intimately related to the processes used to generate forecasts, not only to the algorithmic methods. In the past 25 years, forecasters have tried a number of ways to improve accuracy within a company's forecasting process, from structured judgmental adjustments and statistical forecasts (Armstrong, 2001) to collaborative planning, forecasting, and replenishment (CPFR) along the supply chain (Seifert, 2002). Yet the published surveys on forecast accuracy do not differentiate between respondents based on the maturity of their processes, whether a full-fledged CPFR effort or a part-time employee with a spreadsheet.

Benchmarking is deeply connected to process improvement (Camp, 1989). The two are, in a sense, inseparable. It follows that, as long as information on forecasting processes is not available, we really do not know whether reported MAPEs are "good" or "bad." Forecasting is an art that depends on good methods/algorithms and on sophisticated processes. Using results from purely scientific (what could be called in vitro or lab-based) forecasting competitions such as the M-Competitions or the recent competitions on Neural Network forecasting as benchmarks (Bunn and Taylor, 2001) will be difficult, as these competitions are often dissociated from the processes of the company that provided the data.

Business Model. The published surveys of forecast accuracy have examined business-to-consumer (B2C) sales in retail. In retail, we can only observe sales, not demand—if customers do not find the desired product on the shelf, they will simply shop elsewhere, and the store manager will usually be unaware of the lost sale. The information basis on which a forecast can be calculated is therefore reduced. We may want to forecast demand but only be able to observe historical sales.

This so-called censoring problem is especially serious for products where the supply cannot be altered in the short run, such as fresh strawberries. We may have a wonderful forecast for customer demand but miss sales by a large margin, simply because the stock was not high enough. Thus, comparing the accuracy of a strawberry sales forecast with a napkin sales forecast will be inappropriate: The censoring problems are more serious for strawberries than for napkins.

By contrast, in a business-to-business (B2B) environment, we often know the historical orders of our business clients, so even if the demand cannot be satisfied, we at least know how high it was. Therefore, B2B forecasts profit from much better historical data and should be more accurate than B2C forecasts.

Any published benchmarks on forecasts for products that could be sold either B2B or B2C are consequently harder to interpret than forecasts for "pure" B2B or B2C products.

Moreover, in a build-to-order situation, one may not even know the specific end-products that will be sold in the future. Here it makes sense to either forecast on a component level or to forecast sales volume in dollars rather than in units.

To summarize, none of the published sales forecasting studies can be used as a benchmark. All published indicators suffer from serious shortcomings regarding comparability of data and processes in which forecasts are embedded, as each industry and each company faces its own forecasting problems with its distinctive time granularity, product mix, and forecasting processes. The issues of incomparability have been recognized for many years (Bunn and Taylor, 2001) but have not been solved.

All studies published to date have averaged sales forecasts calculated on widely varying bases, used poorly defined market categories, and ignored the underlying forecast processes at work. These shortcomings are so severe that, in my opinion, published indicators of forecast accuracy can only serve as a very rudimentary first approximation to real benchmarks. One cannot simply take industry-specific forecasting errors as benchmarks and targets.

External vs. Internal Benchmarks

Are the survey problems of comparability resolvable? Could we, in principle, collect more or better data and create "real" benchmarks in forecasting?

The differences between companies and products are so large that useful comparisons among companies within the same market may be difficult to impossible. For instance, even in the relatively homogeneous field of grocery-store sales forecasting, I have seen "normal" errors for different companies varying between 20% and 60% (MAPE for 1-week-ahead weekly sales forecasts), depending on the number of fast sellers, the presence of promotional activities or price changes, the amount of fresh produce (always hard to forecast), data quality, etc. Thus, comparability between different categories and different companies is a major stumbling block.

In addition, industries differ sharply on how much information they are willing to provide to outsiders. I have worked with retailers who threatened legal action if my company disclosed that they were considering implementing an automated replenishment system. These retailers considered their forecasting and replenishment processes as so much a part of their competitive edge that there was no possibility of publishing and comparing their processes, even anonymously. It simply was not to be done. This problem is endemic in the retail market and makes benchmarking very difficult. It may be less prevalent in other markets, but it is still a problem.

My conclusion is that the quest for external forecasting benchmarks is futile.

So what should a forecaster look at to assess forecasting performance and whether it can be improved? I believe that benchmarking should be driven not by external accuracy targets but by knowledge about what constitutes good forecasting practices, independent of the specific product to be forecast.

The article by Moon, Mentzer, and Smith (2003) on conducting a sales forecasting audit and the commentaries that follow it serve as a good starting point to critically assess a company's forecasting practices and managerial environment. It's important to note that no one—not the authors of the paper, not the commentators, and none of the other works made reference to—recommended that you rely on or even utilize external forecast accuracy benchmarks. When discussing the "should-be" target state of an optimized forecasting process, they express the target in qualitative, process-oriented terms, not in terms of a MAPE to be achieved. Such a process-driven forecast improvement methodology also helps us focus our attention on the processes to be changed, instead of the possibly elusive goal of achieving a particular MAPE.

Forecast accuracy improvements due to process and organizational changes should be monitored over time. To support the monitoring task, one should carefully select KPIs that mirror the actual challenges faced by the organization. And historical forecasts as well as sales must be stored, so that you can answer the question, "How good were our forecasts for 2008 that were made in January of that year?" We can then evaluate whether, and by how much, forecasts improved as a result of an audit, a change in algorithms, the introduction of a dedicated forecasting team, or some other improvement project.

In summation, published reports of forecast accuracy are too unreliable to be used as benchmarks, and this situation is unlikely to change. Rather than look to external benchmarks, we should critically examine our internal forecast processes and organizational environment. If we focus on process improvement, forecast accuracy and the use an organization makes of the forecasts will eventually be improved.

REFERENCES

Armstrong, J. S. (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. New York: Springer.

Boylan, J. (2007). Key assumptions in calculating the cost of forecast error. *Foresight: International Journal of Applied Forecasting* 8, 22–24.

Bunn, D. W., and J. W. Taylor (2001). Setting accuracy targets for short-term judgemental sales forecasting. *International Journal of Forecasting* 17, 159–169.

Camp, R. C. (1989). *Benchmarking: The Search for Industry Best Practices That Lead to Superior Performance.* Milwaukee, WI: ASQC Quality Press.

Catt, P. (2007). Assessing the cost of forecast error: A practical example. *Foresight: International Journal of Applied Forecasting* 7, 5–10.

Chatfield, C., M. Hibon, M. Lawrence, T. C. Mills, J. K. Ord, P. A. Geriner, D. Reilly, R. Winkel, and S. Makridakis (1993). A commentary on the M2-Competition. *International Journal of Forecasting* 9, 23–29.

Jain, C. L. (2007). Benchmarking forecast errors. *Journal of Business Forecasting* 26(4), 19–23.

Jain, C. L., and J. Malehorn. (2006). *Benchmarking Forecasting Practices: A Guide to Improving Forecasting Performance,* 3rd ed. Flushing, NY: Graceway.

Kolassa, S., and W. Schütz (2007). Advantages of the MAD/MEAN ratio over the MAPE. *Foresight: International Journal of Applied Forecasting* 6, 40–43.

Makridakis, S., C. Chatfield, M. Hibon, M. Lawrence, T. Mills, K. Ord, and L. F. Simmons (1993). The M2-Competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting* 9, 5–22.

McCarthy, T. M., D. F. Davis, S. L. Golicic, and J. T. Mentzer (2006). The evolution of sales forecasting management: A 20-year longitudinal study of forecasting practice. *Journal of Forecasting* 25, 303–324.

Moon, M. A., J. T. Mentzer, and C. D. Smith (2003). Conducting a sales forecasting audit (with commentaries). *International Journal of Forecasting* 19, 5–42.

Pearson, R. (2007). An expanded prediction-realization diagram for assessing forecast errors. *Foresight: International Journal of Applied Forecasting* 7, 11–16.

Seifert, D. (2002). *Collaborative Planning, Forecasting and Replenishment*. Bonn, Germany: Galileo.

Valentin, L. (2007). Use scaled errors instead of percentage errors in forecast evaluations. *Foresight: International Journal of Applied Forecasting* 7, 17–22.

1.8 DEFINING "DEMAND" FOR DEMAND FORECASTING*

Michael Gilliland

Demand forecasting is often uncritically based on histories of orders received, shipments/sales, or some combination of the two. But as Michael Gilliland explains in this article, the ultimate goal—a measurement of unconstrained true demand—is elusive and not always amenable to simple formulae based on orders and shipments.

Since true demand is not directly measurable, it must be forecast using an approximation constructed from the data we do have available (orders, shipments, backorders, etc.). There is a general belief that orders provide an upper bound to true demand, while shipments (or sales) provide a lower bound, but this is far too simplistic. The relationship depends on reactions to a failure to fill demand in the desired time frame.

So what to do? Recognizing the measurement difficulties, Gilliland suggests we can often derive a proxy for true demand that is "close enough" to be useful in generating an unconstrained forecast. Then, through sales and operations planning or other internal processes, the unconstrained forecast is merged with production/procurement capabilities and inventory availability to generate the "constrained forecast."

* This article originally appeared in *Foresight: The International Journal of Applied Forecasting* (Summer 2010), and appears here courtesy of the International Institute of Forecasters.

An important point is that forecasting performance evaluations should be based on the constrained forecasts, those that represent the organization's best guess at what is really going to happen after taking supply limitations into consideration. We can reliably measure the accuracy of the constrained forecast by comparing it to what really does happen (shipments, sales, or services provided).

Introduction: Unconstrained vs. Constrained Demand

Companies commonly characterize demand as "what the customers want, and when they want it," sometimes with the added proviso, "at a price they are willing to pay, along with any other products they want at that time." When businesses refer to demand, they mean unconstrained or true demand, which does not take into account their ability to fulfill demand. True demand is largely unobservable; so, as a practical matter, we can only approximate it with measurable quantities.

In contrast, the term constrained demand refers to how much demand can be fulfilled in light of limitations on the provision of the product or service demanded. Thus, constrained demand ≤ true demand.

A good forecast of demand, far enough into the future, allows an organization to invest in the facilities, equipment, materials, and staffing required to most profitably fulfill that demand. The planning process begins by loading demand histories into our forecasting software, with the purpose of creating an unconstrained demand forecast. Here, we encounter a problem: What is our operational definition of *demand?* What is the specific, systematic way we measure it?

A company needs to know how to measure true demand in order to provide the proper history for its forecasting models. Typically, you know your orders, shipments, and sales. You know calls handled at call centers, transactions processed at retail stores, and hours billed by consultants. You can track inventory, out-of-stocks, fill rates, back-orders, and cancellations. Still, while you have all these data, none yields the exact true demand.

Orders vs. True Demand

If customers place orders to express their "demand," and if the manufacturer services its customers perfectly by filling all orders in full and on time, then we have our operational definition. In this case, Demand = Orders = Shipments. If both order and shipment data are readily available in the company's system, then we have the historical demand data that we can use to feed our statistical forecasting models.

Unfortunately, few organizations service their customers perfectly—in other words, have an order-fill rate of 100%—so orders are not a perfect indicator of true demand. When some orders received cannot be filled in the customer's desired time frame, several different outcomes are possible:

1. An order that cannot be filled may be rejected by the company or canceled by the customer.
2. An unfilled order may be rolled ahead into a future time bucket.
3. If customers anticipate a shortage, they may inflate their orders to capture a larger share of an allocation.
4. If customers anticipate a shortage, they may withhold their orders, change the orders to a different product, or redirect their orders to an alternative supplier.

In the first case, the cancelled or rejected order may not appear in the demand history file. The omission means that current-period orders will understate true demand.

In the second case, the rolled-ahead order appears in a time bucket later than when it was placed by the customer, so true demand is overstated in future time buckets. That is, the order appears both in the original time bucket and again in future time buckets until the demand is filled or the order is cancelled.

In the third case, a savvy customer (or sales rep), anticipating that product scarcity will lead the supplier to impose an allocation formula (such as "fill all orders at 50%"), will now inflate the order—to twice the true demand, for example.

The fourth case, of withheld or redirected orders, is particularly harmful. Now the historical orders for the desired product do not include the withheld orders, once again understating true demand. Customers may truly want your product, but demand won't be reflected in your historical data because no order was placed. Worse, if customers order a product other than the one they really wanted because of a shortage of the original product, orders for this "second-best" or substitute product overestimate the true demand for the substitute product.

Finally, in a period of chronic supply shortages (due either to supply problems or demand much higher than anticipated), customers may go elsewhere, and all information on their demand is lost.

The assumption is often made that orders provide an upper bound (i.e., will be equal to or greater than true demand), but the four cases noted here reveal that there is no simple arithmetical connection between orders and true demand. In cases 1 and 4, orders will understate demand; in cases 2 and 3 (and sometimes 4, too), orders will overstate demand.

Shipments and Sales vs. True Demand

As with orders, there are also problems in using shipments to represent demand. Shipments are often perceived as a lower limit to true demand; that is, less than or equal to true demand. Thus, shipments and orders are thought to represent true demand's lower and upper bounds, respectively (e.g., Chockalingam, 2009).

We have noted above that cases 1 and 4 show that orders can understate true demand. Furthermore, in case 2, shipments can exceed true demand. This occurs when an unfilled order is rolled ahead into a future time bucket and then filled; the shipment then exceeds true demand in the time bucket in which it is finally shipped. Similarly, in case 4, shipments of a "second-best" product overstate the true demand for the substitute.

Seeking an Operational Definition of True Demand

More complex—but not necessarily better—operational definitions of true demand can be constructed by some hybrid of orders and shipments. Examples include:

1. Demand = (Shipments + Orders) / 2
2. Demand = Shipments + Incremental shortages
3. Demand = Shipments + Latest shortages

The first formula defines demand as halfway between orders and shipments. If order is 120 and shipment is 100, then demand = 110. It simply "splits the difference" by assuming half of the shortages represent legitimate demand, while the rest are due to order manipulation or other gamesmanship.

The second formula avoids overcounting repeat shortage rollovers by only adding increases in shortages to shipments. Therefore, if the shortage in time period t is 20, and the shortage in period $t + 1$ is again 20, then demand = shipments for period $t + 1$ (the shortage amount, 20, did not increase from the prior time period). If the shortage in period $t + 2$ is 25, the demand in period $t + 2$ is shipment + 5 (because there was an incremental 5 units of shortages from 20 to 25).

The third formula also avoids overcounting repeat shortages, in this case by including in demand only those shortages still showing at the end of the time bucket. The demand for a month will include all shipments of that month + unfilled orders of the last week only. If, for example, shortages in a four-week month were 10, 20, 40, and 30, the total demand for the month would be shipments + 30 (the last week's shortages). Table 1.9 illustrates various demand definitions over a one-month period (Gilliland, 2003).

As if this weren't complicated enough, most ERP systems save multiple dates for each order. These may include Order Entry Date, Order Promise Date, Revised Promise Date, Actual Shipment Date, and Customer Receipt Date. Even if companies reach a consensus on how to use order and shipment data, choosing among these dates adds another degree of difficulty in defining demand.

Table 1.9

Week	1	2	3	4	Month Total
Orders	50	50	60	60	220
Shipments	50	40	55	40	185
Shortages		10	5	20	35
Incremental shortage		10		15	25
Latest shortage				20	20

1. Demand = (Shipments + Orders) / 2 = (185 + 220) / 2 = 202.5
2. Demand = Shipments + Incremental shortages = (185 + 25) = 210
3. Demand = Shipments + Latest shortages = (185 + 20) = 205

Chockalingam (2009) illustrates two ways of calculating true demand, starting from either observed bookings (orders) or from observed (gross) shipments:

Observed Bookings

– Requested deliveries in the future

– Exaggerated customer orders

= True Demand

Observed (Gross) Shipments

+ Cuts (unfilled orders that are cancelled)

+ Backorders

– Carryovers

= True Demand

However, because of the vagaries of customer orders, these do not yield operational definitions of true demand. For one, we are unlikely to know the extent of "exaggerated customer orders." And the amount of "cuts" is a function of quantity ordered—yet we saw above that orders are not a reliable indicator of true demand.

To summarize, a suitable operational definition of demand may be unique to each organization and may be difficult to construct, given the available data. For a manufacturer, what a customer orders may not be the same as true demand, nor is true demand what the manufacturer actually ships. For a retailer, what is actually sold off the shelves may not be the same as true demand, either. For example, customers may not be able to find what they want in the store (due to out-of-stocks, or poor merchandise presentation and layout), so there is true demand but no recorded sale. In this case, they may buy a substitute product instead, for which we will record a sale, although there was no original demand.

Determining true demand for a service can be equally vexing. I may wish to stay at a bargain-rate hotel, but have to upgrade when my preferred choice is sold out. Or, I may call the cable company to complain about my television reception, only to hang up in frustration while trying to wade through their voice-menu system.

As a practical matter, while we can't know exactly what true demand really is, we can often come close enough to make the concept useful in forecasting and organizational planning. For manufacturers that do a good job at filling orders (say, 98%+), then shipments, orders, and true demand are virtually the same. Likewise, if a retailer's shelves are fully stocked (or nearly so), then point-of-sale data (cash-register receipts) may be an adequate representation of true demand.

Whether we can provide an accurate proxy for true demand or not, the errors made in approximating true demand can pale in comparison with forecast-model errors.

Making heroic efforts to capture a perfect history of true demand is unlikely to result in significantly improved forecasts and is probably not worth the effort.

True vs. Constrained Forecasts

Forecasts of true (unconstrained) demand provide the right starting point for the planning process (for example, see the *S&OP How-To Handbook* by Wallace and Stahl, 2008). The unconstrained forecast gives the supply chain an unfettered prediction of what customers are going to want in the future, allowing the organization to take action to meet this demand. If future demand is predicted to exceed the current available supply, the organization can hire workers or add shifts, build new facilities, or outsource production. Alternatively, the organization can take steps to reduce demand to levels it can fulfill, such as by increasing prices, dropping customers, or eliminating sales channels.

An output of the planning process is the constrained forecast, which accounts for anticipated supply limitations. The constrained forecast typically is not generated with a statistical model within the forecasting software, but is instead determined through the organization's planning process. It indicates the expected shipments, or expected sales, or expected services that will be provided. It represents the organization's best guess at what is really going to happen—what the shipments, sales, or services provided are really going to be.

Any gap between the true and the constrained forecasts is useful information for managing customer service. For example, when a manufacturer anticipates a shortage, customers can be contacted and their demand redirected to a future date (when their demand can be fulfilled) or to alternative products. It is a failure of management to continue the solicitation of orders when it is known in advance that those orders cannot be filled.

Assessing Forecast Accuracy and Making Financial Projections

Since we are unable to measure true demand reliably, we shouldn't base evaluations of forecast accuracy on our attempts to do so. The "true demand" forecast still serves a valuable purpose—as the starting point in the planning process—but any reports of its accuracy are immediately suspect.

Instead, it is appropriate to assess the accuracy of the forecast for constrained demand. This forecast—what we really expect to ship, sell, or service—is evaluated against what really does happen. Unlike the murky measurement of true demand, an organization should be able to measure unambiguously what really does ship or sell, or the amount of services it provides.

It's important that planners recognize the difference between the unconstrained and the constrained forecast. The planning process should always begin with the unconstrained forecast, nebulous though it may be, as this represents the potential opportunity. When future demand appears to exceed future supply, the organization can take steps to increase supply and meet that demand (or decide not to pursue it, or purposely reduce demand). In contrast, the constrained forecast is an outcome of the planning process and records what the organization ultimately expects to ship, sell, or service.

As a final note, financial projections should always be made from the constrained forecast. It makes no sense to project revenues for any unconstrained demand you know in advance you can't fulfill.

REFERENCES

Chockalingam, M. (2009). What is true demand? *Demand Planning Newsletter* (April). http://www.dem&Planning.net/Newsletters/DPnewsletter_april2009.html.

Gilliland, M. (2003). Fundamental issues in business forecasting. *Journal of Business Forecasting* 22 (2) (Summer), 7–13.

Wallace, T., and R. Stahl (2008). *Sales and Operations Planning: The How-To Handbook,* 3rd ed. T. F. Wallace & Co.

Portions of this material originally appeared in Gilliland, M. (2010), *The Business Forecasting Deal: Exposing Myths, Eliminating Bad Practices, Providing Practical Solutions*. Hoboken, NJ: John Wiley & Sons.

1.9 USING FORECASTING TO STEER THE BUSINESS: SIX PRINCIPLES*

Steve Morlidge

Based on his book, *Future Ready: How to Master the Art of Business Forecasting*, Steve Morlidge argues that business forecasting focuses too narrowly on the short-run forecast of a single variable. While helpful to synchronize demand and supply, this focus makes little contribution to the process of steering business performance. Instead, forecasters need to adopt a broader perspective on the role of strategic forecasting, and take a longer-range view on forecasts themselves.

Morlidge's remedy requires that professional forecasters face the challenges of forecasting the complex behavior of economic systems and address the reality that forecasting is not a stand-alone process. Rather, it exists as part of an organizational control system. Actions taken in response to a forecast (such as increased advertising) often invalidate the assumptions on which the forecast was originally based, making accuracy measurement problematical. Forecast errors might be the result of decisions made and actions taken, not a reflection of the quality of the forecasting process.

While appealing to forecasters to look beyond the short term, Morlidge provides six principles as a roadmap to process improvement. Like Charles Re Corr's article earlier in this chapter, these principles characterize what is needed to create a reliable business forecast to guide decision making.

Economic Forecasting Is Broken

"It's awful! How come no one saw it coming?" Queen Elizabeth spoke these words, in November, 2008, after a briefing on the credit crunch at the London School of Economics. The queen's question echoes that of people at all societal levels; economists and economic forecasters in particular do not currently enjoy a high reputation.

It is not just macroeconomic forecasting that is broken. The financial forecasts used by business executives have also proved highly fallible. "The financial crisis has obliterated corporate forecasts," reports *CFO Magazine* (Ryan, 2009); 70% of respondents to their recent survey said that they were unable to see more than one quarter ahead.

The problem is not restricted to bad economic times. According to the Hackett Group, only 18% of senior finance professionals are "highly satisfied" with their forecast process, and no wonder. On average, earnings forecasts are 13% out, knocking about 6% off their share price (EIU, 2007). Since 2005,

* This article originally appeared in *Foresight: The International Journal of Applied Forecasting* (Winter 2010), and appears here courtesy of the International Institute of Forecasters.

the 1,300 companies quoted on the London Stock Exchange have issued an annual average of 400 profit warnings, each resulting in a loss of value of 10% to 20% of market capitalization (Bloom and colleagues, 2009).

A survey recently conducted for KPMG of 540 senior executives (EIU, 2007) found that improved forecasting topped their priority ranking for the next three years. Ability to forecast results also leads the list of internal concerns for CFOs across the globe (Karaian, 2009).

The problems present a massive opportunity. From nearly 30 years' experience in finance, I can confirm that professional forecasting input is rare in the kinds of forecasts that interest CFOs: medium-term estimates of future revenues, earnings, cash flow, etc. At best, sales forecasts are used to inform short-term forecasts; even then, they are frequently adjusted. Industry surveys suggest forecasts are judgmentally adjusted 72% of the time, and a previous *Foresight* article suggests it could be as high as 91% (Goodwin and Fildes, 2007).

The Narrow Focus of the Forecasting Profession

Another take on this situation is that it represents an indictment of the forecasting profession. In writing this article, I was encouraged by *Foresight*'s editor to read the journal's previous issues. I found much to admire, including many things that I would have included in my book had I known about them before. Most of the contributions are engaging, practical, and intellectually rigorous—a very rare combination in business writing. However, I found very few addressing beyond the very short term. This is no fluke; I have seen little evidence elsewhere that the profession routinely contributes to business forecasting much beyond sales in the near term.

Why should the focus of forecasters be so narrow? Costs and profits are important business variables. Their behavior patterns are neither so obvious as to make the task of forecasting trivial, nor so chaotic as to make it futile. No one would argue that applying forecasting expertise to such problems is not worthwhile. Failure to forecast business performance can be catastrophic, and there is a desperate need for a more rigorous, scientific approach to the task.

If forecasting professionals are to make a bigger contribution to the management process, they must face two challenges.

Challenge #1: The Nature of Economic Systems

Traditional forecasting techniques are usually based on the premise that the future can be predicted by understanding the past. There is increasing recognition of the limitations of this approach. David Orrell and Patrick McSharry noted in their recent *Foresight* article (Orrell and McSharry, 2009), that real-world phenomena such as biological systems, weather patterns, and economic activity are complex and prone to unpredictable behavior, and that is why we can produce reasonable weather forecasts only for a few days ahead, despite decades of huge investment in technology. The phenomena that business forecasters deal with create even greater challenges. There are no economic laws to rival those that we know constrain the behavior of physical objects, and business executives are interested in forecasting beyond the very short term.

Furthermore, little attention is given to forecasting whole systems, as opposed to single variables or multiple instances of a single variable (usually sales or revenue). Steering a business requires forecasts of multiple interdependent variables: volume, price, materials costs, advertising and promotion, infrastructure expense, etc. If professionals shirk this task, then amateurs (such as accountants) will take it on.

A cause for optimism, however, is the growing awareness that discontinuities in systems behavior are not simply inconvenient blemishes in the data record, but matters of vital importance to forecasters and managers. Top executives are more interested in the message "Something is about to change/has changed" than they are in "Everything is as expected." So forecasters should try to anticipate these changes (Batchelor, 2009); if we cannot, we need to get better at spotting them quickly.

Challenge #2: The Organizational System

More attention needs to be paid to the organizational context in which forecasting activities sit. There are excellent articles about the impact of corporate politics on the integrity of forecasts (Finney and Joseph, 2009; Wallace and Stahl, 2009). Many others refer to the purpose of forecasting as informing decision making, but I see no references to the logical corollary that forecasting exists as part of an organizational control system, not as a standalone process. This means that forecasters have to comprehend and contribute to the ways in which targets are set and decisions made, rather than treating these as givens to be managed.

Viewing forecasting as part of a control process has methodological implications too. Take, for example, a fashion retailer. To steer business performance over the course of a season, managers have a range of levers at their disposal: prices, product range, advertising and promotions, store layout, opening hours, staff numbers, and so on. In response to a forecast, management may start, stop, bring forward, put back, and change planned initiatives or create new ones, and this process will repeat monthly or even weekly.

These decisions are taken in pursuit of some clearly stated corporate objective, in response to the actions of competitors, or to exploit competitors' perceived weaknesses. In these circumstances, historical patterns of behavior, upon which so much forecasting technique relies, are of limited value in forecasting future outcomes. Indeed, the value of some variables may be entirely the result of management discretion (e.g., the levels of advertising). Additionally, since management's actions taken in response to forecasts often invalidate the assumptions on which the forecasts were based, measuring forecast accuracy can be problematical. Forecast errors are likely to be the result of decisions taken, rather than a reflection of the quality of the forecast process, and the complexity and dynamism of the situation make it almost impossible to disentangle the impact of one from the other.

How we currently forecast in business is not necessarily wrong or futile. But if they are to make the kind of contribution to businesses they could (and should), forecasters must be more sensitive to the context in which they work, making necessary adjustments to their approach and techniques.

Forecasting is widely used in business to anticipate demand and synchronize with the capacity to fulfill that demand. I see no need for major changes here. Because businesses are designed to be responsive to demand, operational forecasting is short-term in nature. In the short term, systems are unlikely to change trajectory, and decisions made by the business will have limited effect on behavior patterns. The purpose of the short-term forecast is to coordinate an appropriate response to a system that is uncontrollable, just as we use short-term weather forecasts to help us decide what to wear. Consequently, the forecasting methodologies developed over the last few decades—built on the premise that the future will be rather like the past—can work well in these circumstances.

The traditional approach works less well beyond the short term. In the medium-to-long term, managers use forecasts to steer the business and may adopt a different course to what can be anticipated by a forecast based on past data. Also, the longer the time horizon, the greater the chance that the economic system will unpredictably change its behavior because of the nature of the system itself or because of decisions made by other actors, such as competitors. Here, boat racing is a better analogy than weather forecasting. A sailor will continuously reforecast and change course in anticipation of weather conditions, tides, and the movements of competitors.

Much of the opprobrium heaped on forecasters results from failure to recognize the fundamentally different roles that forecasting plays in these two sets of circumstances. The failure is not just of perception. Changes in approach are required if professional forecasters are to make effective contributions to managing business beyond the short term.

Prescription for Change

At a practical level, making these changes means:

- Forecasters have to dispel the notion, in our own and our customers' minds, that forecasting is prophecy. A forecast can be no more than a projection of what might happen, given a set of reasonable assumptions, one of which may be that recent trends will continue. Being able to anticipate outcomes, even imperfectly, means businesses can buy some time to prepare for what might lie ahead. Since the future can adopt any one of a number of trajectories, each of which may demand different responses, this also will involve making not one forecast, but a series of forecasts.

- Forecasters should dispense with the idea that increasing forecast accuracy is their primary aim (Oliva and Watson, 2006). A good medium-term business forecast is one that is accurate enough for the purposes of decision making. In practice, it should be unbiased and with acceptable margin for error. In addition, because we cannot assume the future will be like the past, measures of historic forecast performance have limited value. The focus should be on whether the current forecast is reliable for decision making now, using tools such as Trigg's Tracking Signal (Trigg, 1964).
- Since steering complex, integrated businesses involves a wide range of information and possible responses, forecasters need to handle a large, complex set of interdependent variables and deploy a range of forecasting methodologies, including judgmental forecasting techniques. Having a good understanding of the decision-making processes involved and making pragmatic, well-informed choices about data gathering and modelling is critical if the exercise is not to become hopelessly complicated.
- Forecasting should be perceived not as a stand-alone technical discipline but as part of an organization's performance-management system. This implies requirements for forecasters to have a broader understanding of business process, much more than simply integrating S&OP with the financials.

Another consequence is that there is a much bigger constituency for forecasters to manage. They must ensure that a wider range of business professionals have an appreciation of the fundamentals of good forecasting.

It may sound daunting, but I believe that the job of integrating professional forecasting into the day-to-day business-steering processes can be condensed into six simple principles. I hope these will provide a framework to facilitate the marriage of the skills of professional forecasters with the needs of their customers. Forecasters can use these principles as a map to help them colonize territory beyond the limits of their technical expertise. They also help general managers understand, utilize, and cultivate good forecasting practice.

Forecasting to Steer the Business: Six Principles

Principle 1: Mastering Purpose

Business forecasting is like navigation at sea:

> *It makes sense to plan before a journey, but the original plan (or budget) is often soon outdated because of changes in the weather or tides. Then you need to forecast where you are headed, so that you can determine the necessary corrective action to get to your destination.*

The first thing this example shows is that it's important to distinguish between a forecast (where you think you will be) and a target (where you want to be). Often there is a gap between a forecast and a target, at least until appropriate corrective actions (decisions) are taken—even then, except in the most stable environment, gaps are likely to open up again quickly.

It also helps us to understand the set of qualities needed in a good forecast.

- A good forecast is timely. If you are heading into trouble, a rough-and-ready forecast delivered quickly is much more valuable than a perfect one arriving too late for corrective action.
- It should be actionable. Do we need to make lean on the tiller? Hoist different sails? In business, this means that you need detail only if it is relevant to decision making. You'll probably require different information for forecasting than that used for budgeting. Much more information may be needed about "projects" (e.g., the impact of a new product launch); much less detailed information for "business as usual" (e.g., overhead costs).
- It should also be reliable. As I noted above, a forecast needn't be precise to be reliable; it has to be accurate enough for the purposes of decision making. In practice, it should be free from bias and with acceptable variation.
- It should be aligned. It would be no fun in a storm if every crew member had a different view of where the ship was heading and what course to steer in order to avoid the rocks, yet many businesses have competing "versions of the truth" produced by different functions. In these circumstances, decisions can become driven by corporate politics, slow, and fraught with risk.
- Finally, a forecast should be cost-effective.

Principle 2: Mastering Time

Time is critical to designing and running a forecast process. If a business had perfect information and could react instantaneously, forecasting would be unnecessary. Since this is not the case, two questions must be asked.

First, how far ahead do you need to forecast? The answer depends on how long it takes to enact a decision.

In this example, a supertanker needs to be able to see three miles ahead at all times, since it takes that long to stop. A speedboat, in contrast, requires much less forward visibility. In practice, this means that businesses need a rolling forecast horizon, based on the lead times associated with "steering actions." If an important steering decision—say, launching a new product—takes 12 months, then the business always needs 12 months' forward view. A traditional year-end financial forecast—where the forecast horizon declines the closer you get to year's end—is like overtaking on a blind bend. You have no idea of the possible outcome of your decision.

How frequently should you forecast? That depends on how quickly things change. A ship's captain needs to forecast more frequently in the busy Singapore Strait than she does in the wide-open spaces of the South Pacific. Accounting-period ends should not determine the timing of forecasts. Southwest Airlines, for instance, updates revenue forecasts daily, but aircraft-ownership costs (leases, depreciation, etc.) only once quarterly.

Principle 3: Mastering Models

Any forecast requires a model, a set of assumptions about the way the world works. This could be a statistical model, one extrapolating into the future from the past. Alternatively, it may be a causal factor model, based on the identification of key drivers. If the future is like the past, these kinds of models can be very effective.

However, often the world is too complex or the business is changing too fast to make such well-structured approaches workable. That is why business forecasting frequently relies on judgment: where the model is in the head of an expert or a larger number of people who "understand the business" or "know what is going on in the market." But human judgment has flaws, and managers can feel pressure to adjust forecasts to avoid "nasty surprises" or "sounding defeatist." Consequently, judgmental forecasts are prone to bias.

With forecast models, the trick is to understand the range of methodologies available, choose appropriately, then take steps to mitigate weaknesses. For example, it might be appropriate to use a statistical model to produce a baseline or "business as usual" volume forecast, and judgment to estimate the impact of decisions that alter the course of affairs (e.g., price changes).

Principle 4: Mastering Measurement

If you rely on a forecast to make decisions, the way that forecast is generated should have been reliable in the past. Yet, few businesses take the simple steps required to monitor their processes for evidence of bias so they can take action to eliminate it, if detected. As Jim Hoover notes in his article on the tracking of forecast accuracy (Hoover, 2009), most businesses fail to track forecast quality over time.

Those businesses that do attempt to monitor it often measure the wrong things at the wrong time. Forecast error needs to be measured over the short term—before decisions informed by the forecast have taken effect. To do otherwise is like blaming the navigator for having forecast a calamity that never occurred, precisely because the captain acted upon the forecast and changed course.

Also, forecasts need to be made frequently because it is important to distinguish between inevitable unsystematic error (variation) and systematic error (bias). A sequence of four errors with the same sign (positive or negative) is needed to be able to distinguish bias from the effects of chance. The common business practice of using quarterly forecasts to steer toward an annual target makes it impossible for managers to identify and correct a biased forecast in time.

Principle 5: Mastering Risk

Our only absolute certainty about the future is that any forecasts are likely to be wrong. Debate about the forecast should not focus on whether you have the right "single-point forecast," but how it might be wrong, why, and what to do about it. In particular, it is important to distinguish between *risk*—random variation around a realistic single-point forecast—and *uncertainty* resulting from a shift in the behavior of a system that invalidates the forecast.

A major contributor to the economic collapse earning the queen's comment was the overreliance of banks on risk models that failed to take account of some important sources of uncertainty. Whatever form your ignorance of the future takes, it is important to develop the capability to spot and diagnose deviations from forecast quickly and to create a playbook of potential actions enabling swift, effective response.

Principle 6: Mastering Process

Forecasting is neither art nor complex science. It is mainly a matter of applying modest amounts of knowledge, in a disciplined and organized fashion, as a process. A good process—like a good golf swing—produces good results.

Building a good process involves doing the right things in the right order (cultivating a good technique), over and over (grooving the swing). Those things that are responsible for bias (hooks and slices) should be designed out of the process (remodeling the swing), the results of the process continuously monitored (the score), and minor flaws corrected as they become evident. Again, as in golf, temperament is as important as technique. Blaming people for failures when the process is at fault is a sure way to encourage dishonest forecasting.

Efforts to improve forecasting processes can be undermined by behaviors associated with adjacent business processes such as traditional budgeting.

1. Budgeting does not recognize the distinction between a target and a forecast. Bias in forecasting is often associated with the desire not to show gaps between the two, either because a shortfall is interpreted as poor performance or a "lack of commitment" or because submitting an over-target forecast triggers a target increase.
2. Budgeting is incompatible with the need for rolling horizons built around decision-making lead times—everything is pegged to the financial year-end.
3. By fixing budgets on an arbitrary annual cycle, budgeting constrains an organization's ability to respond and therefore undermines the value of forecasting. Often appropriate, timely action cannot be taken because the department involved "hasn't got the budget."

Conclusions

There is an urgent need to improve the quality of the forecasts used to steer businesses, which hitherto have suffered from a lack of input from the professional forecasting community. To make the kind of contributions that businesses need, there must be a shift in understanding the role of forecasting and a complementary change in how forecasting methodologies are applied. Both require improved relationships between forecasting specialists and their business constituency.

The six principles of forecast mastery noted here help these two parties develop shared understanding of what it takes to create a reliable business forecast. Specifically, there must be the recognition:

1. That the PURPOSE of forecasting is to guide decision making, not to prophesy an outcome.
2. That TIMELINESS is important in managing the decision-making process.
3. That many different types of MODELS can be used to forecast, and that no one technique can be a "silver bullet."
4. That careful MEASUREMENT is needed to assess the reliability of forecasts.
5. That considerations of RISK should not be excluded from the forecast process.
6. That forecasting is part of a disciplined, collaborative, performance-management PROCESS.

REFERENCES

Batchelor, R. (2009). Forecasting sharp changes. *Foresight: International Journal of Applied Forecasting* 13 (Spring), 7–12.

Bloom, A., Wollaston, A. and McGregor, K. (2009). Analysis of profit warnings. Ernst and Young.

EIU (2007). *Forecasting with Confidence: Insights from Leading Finance Functions*. KPMG.

Finney, A., and M. Joseph (2009). The forecasting mantra: A holistic approach to forecasting and planning. *Foresight: International Journal of Applied Forecasting* 12 (Winter), 5–14.

Goodwin, P., and R. Fildes (2007). Good and bad judgment in forecasting: Lessons from four companies. *Foresight: International Journal of Applied Forecasting* 8 (Fall), 5–10.

Hoover, J. H. (2009). How to track forecast accuracy to guide forecast process improvement. *Foresight: The International Journal of Applied Forecasting* 14 (Summer), 17–23.

Karaian, J. (2009). By the numbers: Top ten concerns of CFOs. CFO Europe, September 2009, 11–12.

Oliva, R., and N. Watson (2006). Managing functional biases in organizational forecasts. *Foresight: International Journal of Applied Forecasting* 5 (Fall), 27–32.

Orrell, D., and P. McSharry (2009). A systems approach to forecasting. *Foresight: The International Journal of Applied Forecasting* 14 (Summer), 25–31.

Ryan, V. (2009). Future tense. *CFO* http://www.cfo.com/printable/article. cfm/12668080.

Trigg, D. W. (1964). Monitoring a forecasting system. *Operational Research Quarterly* 15, 271–274.

Various (2008). Aligning forecasting practice with market dynamics. The Hackett Group, Volume 12, Number 1.

Wallace, T., and B. Stahl (2009). Sales forecasting: Improving cooperation between the demand people and the supply people. *Foresight: International Journal of Applied Forecasting* 12 (Winter), 14–20.

1.10 THE BEAUTY OF FORECASTING*

David Orrell

Nobel economist Paul Krugman wrote in 2009 that "[t]he economics profession went astray because economists, as a group, mistook beauty, clad in impressive-looking mathematics, for truth." David Orrell, author of *Truth or Beauty: Science and the Quest for Order* (2012), now asks whether this same hubris has applied to forecasters in general.

Newton's law of gravity is held as the archetype of a beautiful theory—and a predictive model. It possesses the three key aesthetic properties of elegance, unity, and symmetry, and can accurately predict a broad range of phenomena. Orrell compares this to the widely used technique for estimating the risk for any financial asset, known as value at risk (VaR). While

* This article originally appeared in *Foresight: The International Journal of Applied Forecasting* (Winter 2014), and appears here courtesy of the International Institute of Forecasters.

VaR is mathematically elegant, incorporates assumptions of symmetry and stability, and unifies the description of a broad range of phenomena, it has failed on a regular basis. As Orrell tersely puts it, Newton's laws got us to the moon, and VaR got us to the financial crisis.

Citing empirical tests such as the M3 competition, simple models are often better at making predictions than more complicated models. But there is a distinction between simple models that involve few parameters (such as a random walk, or single exponential smoothing), and models like VaR that incorporate assumptions that were overly influenced by theoretical criteria such as symmetry or equilibrium.

Orrell concludes that because "living systems . . . resist the tidiness of mathematical laws," it is a risky business indeed to assume that these systems we seek to analyze are either easily depicted or predictable through elegant equations. So when it comes to predictive models, maybe it's OK if they are a little ugly.

Introduction

Most business forecasters would not associate their field with a quest for beauty. Excel spreadsheets are not renowned for their attractiveness. No one—even, I daresay, its inventors—would claim that a useful tool such as exponential smoothing, or even the "autoregressive integrated moving average," is the most beautiful formula ever devised.

But a sense of aesthetics plays an important, if subtle, role in many branches of science. Bertrand Russell wrote, "Mathematics, rightly viewed, possesses not only truth, but supreme beauty—a beauty cold and austere, like that of sculpture." The same kind of beauty is sought and appreciated by researchers in more applied areas as well—not just for its own sake, but because it often seems to indicate that one is on the right path.

The British physicist Paul Dirac went even further, arguing that "It is more important to have beauty in one's equations than to have them fit experiment." He demonstrated this by using an elegant equation to infer the existence of antimatter before it had been physically detected. Modern "theories of everything" such as supersymmetry are explicitly based on aesthetic ideas (e.g., lots of symmetry), though thus far to much less success.

Forecasters might not go to such extremes—they would not predict a recession or boom just because it "looked good"—but the models they use often carry in their bones a trace of mathematical elegance, which can either help or hinder their accuracy.

Perfect Model

Three key aesthetic properties are elegance, unity, and symmetry. Perhaps the archetype of a beautiful theory—and a predictive model—is Newton's law of gravity. The equation is mathematically simple and elegant. It unifies a broad range of phenomena—everything from the motion of the moon around the earth, to an apple falling to the ground. And it is highly symmetric, both spatially (it is the same in every direction) and in the sense that it produces a symmetric force (the earth pulls on the moon, but the moon also pulls back, causing tides). Physicists seek out symmetries in a system because these allow simplified mathematical representations that can be used to predict the system's behavior.

The success of this reductionist approach set a standard for other fields of science, including economics. Neoclassical economics was founded in the 19th century by economists such as William Stanley Jevons and Leon Walras, who took their inspiration directly from Newton. One of the great appeals of their theory was the physics-like way in which it reduced a complex world to a set of elegant equations. As Jevons put it in his 1871 book, *Theory of Political Economy*, these laws were to be considered "as sure and demonstrative as that of kinematics or statics, nay, almost as self-evident as are the elements of Euclid, when the real meaning of the formulae is fully seized."

The theory's key planks include rationality, stability, and uniformity (a large number of consumers and producers with similar characteristics), which together impose a kind of symmetry on the system. Rationality is a symmetry, because rational people with identical preferences will make the same decision given the same information. Stability is symmetry-in-time—if markets are in equilibrium, then the future looks like the past. And if markets are uniform in the sense that market participants have similar power and other characteristics, then that means transactions are symmetric.

Of course, no one thinks that people are perfectly rational, or that markets are perfectly stable or uniform, and much work has been done exploring deviations from these assumptions. But when it comes to what Krugman called the "impressive-looking mathematics" used in mainstream economic models, the world is a very rational, stable, and uniform place. The same hubris applies to certain forecasting models.

Economy at Risk

For example, policy makers often rely on economic predictions made using general equilibrium models. As the name suggests, these explicitly assume the existence of an underlying market equilibrium, and attempt to simulate how it will rationally adapt to changing conditions. Risk models used by banks also usually assume rational behavior and an underlying equilibrium.

Consider the development of the widely used technique known as value at risk (VaR), which is supposed to estimate the worst-case loss that an institution could face on a given financial position. Risk is calculated by taking historical data over a time window ranging from a few months to several years, depending on the case, and applying standard statistical techniques to give the likelihood of a particular loss in the future.

The model is based on the idea that prices are drawn to a stable equilibrium, but are perturbed randomly by the actions of independent investors or by unexpected news. These assumptions justify the use of elegant statistical methods such as the normal distribution. The risk of an asset can be reduced to a single number based on its historical variation.

Despite its popularity, and its intellectual attractiveness, the model has failed on a regular basis. In 2007, for example, the CFO of Goldman Sachs complained that they "were seeing things that were 25-standard-deviation moves, several days in a row." A 25-standard-deviation event is something that is not expected to happen even once in the duration of the universe—so if it happens several days in a row, you begin to realize there is a problem. In fact, market fluctuations do not follow a normal distribution. Like earthquakes, they are better described by a power-law distribution, which has "long tails" and therefore greater likelihood of extreme events.

Value at risk is certainly mathematically elegant. Just as Newton's equation can describe a broad range of phenomena, so VaR can give an estimate of risk for any financial asset. It incorporates assumptions of symmetry and stability. And like a financial law of gravity, it appears to make the trajectory of markets reassuringly rational and predictable. Unfortunately, it lacks empirical validity. Newton's laws of motion got us to the moon; VaR got us to the financial crisis.

Lessons from Business Forecasting

Empirical tests such as the M3 competition, described in Morlidge (2014), have often shown that simple models are better at making predictions than more complicated models. However, there is a distinction between empirical models that involve few parameters, and models that, like VaR, incorporate assumptions that were themselves overly influenced by theoretical criteria such as symmetry or equilibrium.

As I argue in *Truth or Beauty: Science and the Quest for Order* (Orrell, 2012), a concern with aesthetics has affected our choice of models in many areas of science, from string theory to weather prediction. Of course, not all forecasting or risk assessment tools suffer from the same drawbacks. In fact, I believe the business forecasting community has much to teach other fields about adopting an approach that is pragmatic and realistic.

Consider for example the traditional approach to climate forecasting. This involves producing a mechanistic model of the climate system, based on quasi-Newtonian equations of fluid flow, supplemented by empirical laws for things such as turbulent flow. As anyone who has worked with weather or climate models knows, they are not the most elegant of mathematical constructions;

however, they are still based on a reductionist approach which assumes that, in principle, the behavior of a system's components—and therefore the system itself—can be predicted using simple equations.

An alternative to the mechanistic method is to take a time-series modeling approach. For example, neural network models set up a network of artificial "neurons" that learn to detect patterns in past data. A recent study (Fildes and Kourentzes, 2011) showed that, for a limited set of historical data, a neural-network model outperformed a conventional climate model, while a combination of a time-series model with a conventional model led to an improvement of 18 percent in forecast accuracy over a 10-year period. Such time-series models are particularly good at spotting local variations, which tend to elude traditional models but are very relevant for policy makers.

So why have these statistical time-series techniques not been incorporated for use in official climate forecasts? I would venture that part of the reason is related to aesthetics—a topic that I doubt comes up much at meetings of the Intergovernmental Panel on Climate Change. Methods such as neural networks are based on a set of equations, but if you write them out they seem haphazard and strange. Instead of looking like good, mechanistic science, they look like a hack job.

But perhaps that awkwardness is just an expression of the fact that the system under analysis is not easily reconciled with simple equations. Living systems—such as a cell, a person, an economy, or even the climate (which is produced by life)—resist the tidiness of mathematical laws.

When it comes to predictive models, maybe it's OK if they are a little ugly. After all, as in life, looks don't count for everything.

REFERENCES

Fildes, R., and N. Kourentzes (2011). Validation and forecasting accuracy in models of climate change. *International Journal of Forecasting* 27 (4) (Oct–Dec), 968–995.

Morlidge, S. (2014). Do forecasting methods reduce avoidable error? Evidence from forecasting competitions. *Foresight: International Journal of Applied Forecasting* 32 (Winter), 34–39.

Orrell, D. (2012). *Truth or Beauty: Science and the Quest for Order*. New Haven, CT: Yale University Press.