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

Stephan Kolassa

PREVIEW

Organizations often seek *benchmarks* to judge the success of their forecasts. Reliable benchmarks would allow the company or agency to see if it has improved upon industry standards and to evaluate whether investment of additional resources in forecasting would be money well spent. But can the existing benchmark surveys be trusted? "No," says Stephan Kolassa, who has analyzed the surveys and found them seriously deficient. In this article Stephan explains the many problems that plague benchmark surveys and advises that companies should redirect their search from external to internal benchmarks since the latter provide a better representation of the processes and targets the company has in place.

KEY POINTS

- In *benchmarking*, comparability is the key. Benchmarks can be trusted only if the underlying process to be benchmarked is assessed in similar circumstances.
- Published surveys of forecast accuracy are not suitable as benchmarks because of incomparability in product, process, time frame, granularity, and key performance indicators.
- It is doubtful that forecasting accuracy benchmarks can be compiled from cross-company surveys because the hurdles of establishing comparability are formidable.
- Quantitative targets themselves may be elusive. 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.

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

Stephan Kolassa is Vice President of Corporate Research at SAF AG in Switzerland. He has worked extensively with some of Europe's largest retail chains in producing automatic forecasts for large batches of products. Stephan and his colleague Wolfgang Schütz coauthored "Advantages of the MAD/MEAN Ratio Over the MAPE" in *Foresight*'s Spring 2007 issue.

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

Teresa McCarthy and colleagues (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) 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.

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 2. 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 3.

Table 1. MAPEs for Monthly Sales Forecast in 1984, 1995 and 2006 Surveys

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

Source: McCarthy et al. (2006)

Table 2. MAPEs for Monthly Sales Forecast

Source: Jain & Malehorn (2006, Table 6.2)

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

Tables 2 and 3 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 2), the 1-year-ahead MAPEs are lower than those at shorter horizons.

Unfortunately, the validity of these results is again problematic. The sample sizes were very small in many categories (Table 2), reflecting a low response rate by the attendees. Jain (2007) does not even indicate the number of responses behind the results in Table 3. 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.

Table 3. MAPEs for Monthly Sales Forecast

Source: Jain (2007)

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 & 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%

Table 4. MAPEs for Monthly Sales Forecast

Source: Makridakis et al. (1993)

Company	Industry	Number of series	Forecast	1 month	2 months	1 quarter	1 year
Honeywell	Residential construction	6	Average	N/A	16.6%	15.9%	19.3%
			Best (Naive method including seasonality)	N/A	5.1%	6.7%	13.5%
Squibb	Pharma	7	Average	N/A	9.1%	10.6%	28.1%
			Best (Smoothing with dampened trend)	N/A	7.3%	7.2%	23,0%
Car company	Automotive	6	Average	10.1%	10.7%	14.6%	13.9%
			Best (Smoothing with dampened trend)	8.0%	9.5%	14.6%	14.2%
Aussedat-Rey	Paper	4	Average	3.7%	5.6%	6.8%	5.2%
			Best (Combination of smoothing methods)	2.8%	5.9%	6.7%	3.8%

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 which might influence sales. Table 4 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).

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 4). 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 nonservice 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

Forecasting is an art which depends on good methods/algorithms *and* on sophisticated processes. Using results from purely scientific forecasting competitions will be difficult, as these competitions are often dissociated from the processes of the company that provided the data.

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 which 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 & 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 & 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 upon 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.

CONTACT
Stephan Kolassa
SAF AG, Tägerwilen, Switzerland
stephan.kolassa@saf-ag.com

REFERENCES

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

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

Bunn, D.W. & Taylor, J.W. (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: The International Journal of Applied Forecasting*, Issue 7, 5-10.

Chatfield, C., Hibon, M., Lawrence, M., Mills, T.C., Ord, J.K., Geriner, P.A., Reilly, D., Winkel, R. & Makridakis, S. (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), Winter 2007/2008, 19-23.

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

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

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

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

Moon, M.A., Mentzer, J.T. & Smith, C.D. (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: The International Journal of Applied Forecasting*, Issue 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: The International Journal of Applied Forecasting*, Issue 7, 17-22.

MEASURING IMPROVEMENT IN FORECAST ACCURACY

A CASE STUDY

Robert Rieg

PREVIEW

Over the past 15-20 years, improvements in forecasting methods, deepening practical experience, and increasing computing power should have allowed companies to significantly improve their forecasting accuracy. In this paper Robert Rieg examines the changes in forecasting accuracy of a large automobile manufacturer between 1991 and 2005. His analysis shows how a company can examine its track record over time and emphasizes the need to distinguish internal from external factors that impinge on forecasting accuracy.

IMPROVING FORECASTING ACCURACY OVER TIME

There are four basic means for improving forecast accuracy over time:

(1) Use better forecasting methods/algorithms.
(2) Acquire better software and hardware.
(3) Learn from past experience and mistakes.
(4) Reduce the uncertainty in the forecasting environment.

Deterioration in forecasting performance, or at least an absence of evidence of improvement, could occur despite reasons (1) and (2) because of an increasingly uncertain environment or loss of organizational knowledge. Such a result should prompt deeper analysis of the underlying factors. Are they internal factors, such as change of processes or use of inappropriate methods? If so, the problem is resolvable. However, external factors, such as an increasingly uncertain forecast environment, are considerably more difficult to resolve.

(1) Better forecasting methods/algorithms. The passage of time has seen a considerable enhancement of the toolbox of forecasting methods. But method upgrades do not automatically lead to better predictions. In the M3 forecasting competition (Makridakis & Hibon, 2000) which compared the performance of common forecasting methods on large, diverse data sets, newer, more sophisticated methods like Box-Jenkins and Artificial Neural Networks failed to outperform older and simpler methods such as Exponential Smoothing. Armstong (2006) cites an analysis that data mining, a very complex methodology, fails to improve even upon "random guessing." It is possible that the accuracy gained in upgrading forecast-method selection is initially very high, but the additional returns to increasing sophistication are negligible.

(2) Increased computing power and sophisticated software available at low costs. Companies can now process and store more data with ever-more-complex algorithms in a shorter period of time (Küsters et al., 2006). Some empirical studies show that the use of appropriate software can lay the groundwork for improved forecasts (Sanders & Manrodt, 2003), especially when the organization shifts from paper-based forecasts or spreadsheets to dedicated forecasting software.

(3) Improved learning, training, and knowledge sharing. New methods and software will prove beneficial only if organizations use them in a sensible way. Forecast quality can improve through training, as

Robert Rieg is a Professor of Management Accounting and Control and currently Dean of the Faculty of Business at Aalen University, Germany. He is interested in forecasting for planning, budgeting, and accounting purposes. His latest research project is a panel study on prediction markets.

KEY POINTS

- Improvement over time in an organization's forecasting accuracy can have several sources: better methods, better software and hardware, an improved learning curve, and reduced uncertainty in the organization's environment.
- A major longitudinal study suggested that, over the 20-year period ending in 2006, forecast accuracy had not improved but deteriorated, due partly to over-reliance on forecasting software and failure to appreciate the role of organizational processes and training.
- However, this survey is of dubious validity because it compares the performance of different actors at different times. In place of a longitudinal survey, a case study has promise because it concentrates on one company and allows for application of appropriate metrics for forecasting errors.
- In my case study at a large German automotive manufacturer, I used the metric MAD/Mean to trace changes in forecasting accuracy over the 15- year period, 1991-2005. The MAD/Mean overcomes several major deficiencies of the more traditional metric, the MAPE.
- My results for this company reveal little evidence of improvement over this time period, which I found surprising. It is possible that the forecast environment had become more uncertain over this period, offsetting potential internal improvements. For examinations of forecasting accuracy improvement, it is important to separately identify the effects of internal and external factors.

well as through sharing knowledge about appropriate methods (Byrne & Heavey, 2006). Also important is the establishment of specific organizational units for forecasting, as well as the alignment of forecasting and incentive systems.

(4) Uncertainty and volatility. Forecasting methods need to recognize patterns (e.g trend, seasonality, and structural breaks) and how they change over time. Improved pattern recognition may lead to better forecasts. However, environmental changes may make pattern recognition more difficult by altering historical relationships and by inducing greater volatility. These problems afflict the modeling process across the range, from macro-economic forecasting to forecasting for call centers (Minnucci, 2006). And newly influential variables have to be detected and incorporated into models, increasing the challenges faced by market analysts and forecasters.

The four factors are closely intertwined. Advanced statistical methods are of use only if implemented in software. Better forecasts will lead to better decisions only if organizational processes facilitate the use of additional predictive information. In today's business world factors (1) and (2) should not present a constraint to better forecasting. The more substantial issues concern factors (3) and (4) and which of them prevails.

THE MC CARTHY LONGITUDINAL STUDY

In a review of previous studies supplemented by their own survey of forecasting changes over a 20-year period, Teresa McCarthy and colleagues (McCarthy et al., 2006) found that forecast accuracy had deteriorated over time. They surmised that this grim result was attributable to reduced practitioner familiarity with forecasting methods and to failures in training, processes and performance measurement, and rewards (Category 3 above). They also noted a tendency of managers to rely on forecasting software as a primary solution to their forecasting problems. The concentration on software and under-emphasis on training results in users who don't know what the software does and who tend to accept software results unchecked ("black-box" forecasts).

The McCarthy study reports that only a minority of companies tie compensation incentives to forecast results. Different departments within the company are seldom forced or encouraged to align their different forecasts, a problem that is only recently being addressed through Sales and Operations Planning initiatives.

However, just as Stephan Kolassa concludes (in his preceding article in this issue) that benchmarking is difficult to do from external longitudinal surveys, so the assessment of forecast improvements from such surveys faces the same insurmountable challenges.

McCarthy's surveys:

- Were based on questionnaires. It is hard to control who responds to a questionnaire (key informant bias).
- Used accuracy-metric calculations that seem to have been done by the respondents. We do not know how they did them or how reliable the answers are.
- Compared different studies at different points in time. From comparative-static analyses, one cannot be sure to capture dynamics and trends correctly.
- Included different companies. It is not certain that the responding companies were the same over time, leading to selection biases and survivorship bias.

As an alternative to external surveys, a case study of an individual organization has promise. Using original data avoids informant bias and allows for application of appropriate metrics for forecasting errors. The concentration on one company avoids selection biases. And while the results are not scientifically generalizable, the analysis of a typical manufacturer in a mature industry should be indicative of the entire industry and possibly beyond.

THE GERMAN AUTOMOTIVE MANUFACTURER

Working with a large automobile company, I collected monthly data on **actual** and **planned sales volumes** for three car models sold in six countries. The monthly data span the period 1991 to 2005. The cars are sold in several versions, usually as middle class, upper-middle class, and premium models. The typical life cycle of a version is about seven years. The company has enjoyed decades of successful production and sales of millions of cars. In addition to the collection of sales data, I conducted interviews with company managers to qualitatively assess their forecasting and planning methods.

The company does not differentiate between forecasts (in terms of predicting future events) and plans (in terms of targets for employees). However, the sales target data in this company provide an acceptable proxy for forecasts. Compensation incentives for sales force and sales managers are based on accuracy of achieving planned sales, giving sales personnel motivation to prepare plans close to what they believe will be sold. I did a detailed analysis of plan-actual variances and found no pattern or systematic bias, such as "plans are always higher than actuals."

Given the company's long record of experience, successful market positioning, and large reserve of human and IT resources, one would assume the company would have developed a good forecasting track record. And with data spanning a long period, 15 years, we should be able to detect learning effects.

MEASURING FORECAST ACCURACY IMPROVEMENT

The Metric.

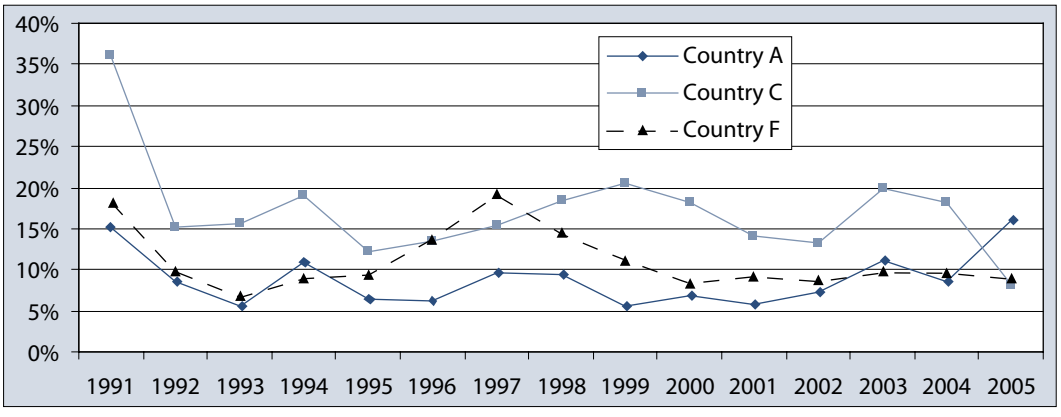
While many forecast error metrics are in use, analysis of forecast-accuracy improvement requires a metric that is not denominated in volume (e.g +/- so many cars), since volumes are very different for different model cars. Additionally, there are months of missing data so that the most common percentage error metrics (MAPEs) cannot be calculated. In their *Foresight* article, Kolassa and Schütz (2007) propose a metric that is unit free and appropriate for interrupted data, the MAD/Mean (the mean average deviation divided by mean sales volume). The metric they show can be interpreted as a weighted average percentage error.

For each of three car models, in each of six countries, we calculated an annualized MAD/Mean ratio over the 15-year period 1991-2005. The annualized figure is the yearly average of monthly forecast errors (actual sales

Table 1. Annualized MAD/Mean for Car Model 1

Years	Country A	Country B	Country C	Country D	Country E	Country F
1991	15.1%	16.6%	36.2%	13.9%	20.8%	18.1%
1992	8.6%	5.8%	15.3%	16.8%	21.1%	9.8%
1993	5.5%	16.5%	15.6%	19.3%	20.7%	6.8%
1994	10.9%	11.7%	19.1%	16.1%	19.4%	9.0%
1995	6.4%	9.6%	12.2%	20.3%	19.8%	9.5%
1996	6.2%	16.6%	13.6%	9.0%	14.3%	13.6%
1997	9.6%	12.1%	15.4%	13.8%	16.1%	19.2%
1998	9.4%	23.7%	18.5%	11.5%	27.3%	14.5%
1999	5.5%	13.6%	20.6%	8.8%	19.7%	11.1%
2000	6.8%	13.2%	18.2%	7.4%	9.6%	8.4%
2001	5.7%	8.1%	14.2%	6.9%	12.8%	9.2%
2002	7.3%	10.0%	13.3%	7.2%	8.6%	8.7%
2003	11.0%	15.6%	19.9%	5.3%	16.0%	9.8%
2004	8.6%	21.7%	18.2%	18.0%	10.6%	9.7%
2005	16.0%	17.3%	8.1%	12.4%	18.8%	9.1%

Figure 1. Annualized MAD/Mean Model 1 for Three Countries, Based on Table 1



– target) divided by the yearly average of monthly sales. The results for the first car model are shown in Table 1 and a portion of this table is plotted in Figure 1.

Detecting Trends in Forecast Accuracy Over Time.

A decreasing trend in forecast errors over time should represent a pattern of improvement in forecasting, while an increasing trend in forecast errors would suggest deterioration. Based on the accuracy metric (MAD/Mean) calculated for different countries and different time periods, we tested for indications of these decreasing and increasing trends. Our test results – applied to all three car models – do not indicate an overall trend towards improved forecasts. The specific statistical test employed is described in the on-line appendix. See www.forecasters.org/foresight/documents/Rieg_Issue11.pdf

You can see in Table 1 and Figure 1 that in some years and countries forecast errors are trending downward, but these are not enduring. Similar results were found for car models 2 and 3. We cannot conclude that there has been an improvement in overall forecast accuracy.

Learning Effects from One Product Life Cycle to the Next.

While evidence of improvement in forecast accuracy over time does not emerge, improvement in forecasts due to learning effects could still be possible from one *product life cycle* (PLC) to the other. Each life cycle is roughly seven years long. In a new life cycle, models with new technology and/or design changes are introduced into the markets while the basic car model stays the same.

For each car model and country, we compared forecast errors that occurred in the first month of two successive life cycles. For example, we compared the MAD/Mean for July 1991, the initial month of PLC1, with that of August 1998, the initial month of PLC2. We repeated the comparison for each of the remaining months. The results are shown in Table 2.

We detected downward changes in forecast errors in only 18% of the comparisons, which was essentially the same frequency of upward changes (deterioration in forecast accuracy). So the majority of comparisons revealed no indications of learning from one life cycle to the next.

Internal Vs. External Factors Affecting Forecast Errors.

As we have noted, changes in forecasting accuracy over time can be attributed to 4 types of factors: changes in methods/algorithms, software and

Table 2. Upward or Downward Change between Life Cycles (Up Implies Deterioration)

Mann/Kendall trend test for subsequent product life cycles (plc)

	Model 1			Model 2		Model 3	
Country	plc 1 to plc 2	plc 2 to plc 3	plc 3 to plc 4	plc 1 to plc 2	plc 2 to plc 3	plc 1 to plc 2	plc 2 to plc 3
A	up	up	(no data)	down	up	down	down
B	up	-	(no data)	-	-		down
C	-	-	-	-	-	-	-
D	up	-	(no data)	-	-	up	-
E	-	-	-	down	-	down	-
F	-	down	-	-	-	up	-

LEGEND	meaning	happens	of all tests
up	upward trend, (α = 5%)	7 times =	18%
down	downward trend, (α = 5%)	7 times =	18%
-	no clear trend	24 times =	63%

hardware, people and organizations, and the forecasting environment. The first three are internal changes while the fourth is an external factor. Perhaps our finding that forecast accuracy failed to improve (or show any trend) over time is attributable to offsets between internal and external events, internal improvements being offset by an increasingly uncertain external environment. One lucid example of this was the unforeseen changes in consumer behavior prompted by environmental concerns about large and polluting German vehicles compared to greener Japanese cars with hybrid engines.

In interviews with company officials, however, I was told the company had not made significant changes in the internal factors – in processes, tools, or algorithms. For a company with a long-standing record of corporate success, vast resources, knowledge, and experience, this was surprising. However, since I relied upon interviews held afterwards, I can't rule out whether there were unreported internal changes or, if so, whether any of these were successful.

Analysis of Change in the Forecasting Environment. Changes in the external forecasting environment evolve gradually over time. One way to detect the magnitude of these changes is to compare the variability – degree of fluctuation – of the sales data at different points in time.

Figure 2 shows the actual monthly sales volume as well as a 12-month moving average for car model 1 in two countries. One can see the 7-year product life cycle and a pattern that shows fluctuations but without evidence that these are increasing or decreasing over time. We do see that some seasonal patterns, such as those for Country B, have smaller peaks after 1998. Such changes in patterns are hard to forecast if one has only time series at hand.

One measure of variability is the standard deviation. For each car model and country, we calculated the annual standard deviations of actual sales volumes and then tested this for decreasing or increasing trends. The results were a mixed picture of upward,

Figure 2. The 7-year Life Cycle for Car Model 1

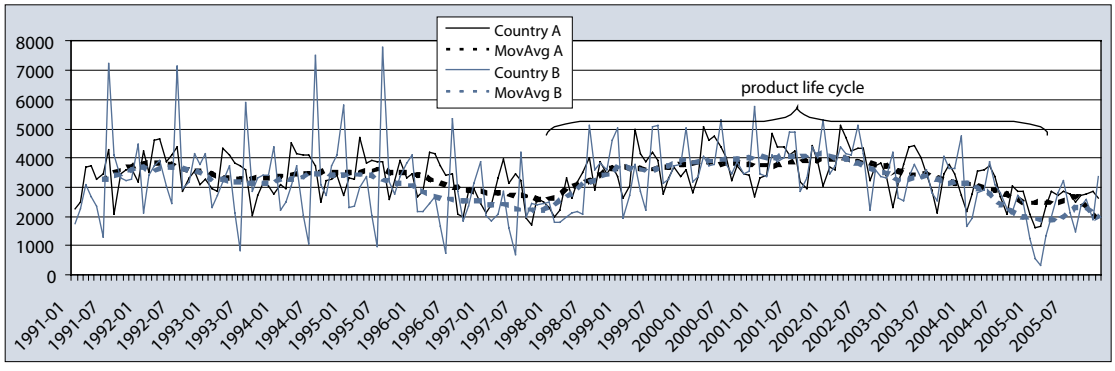
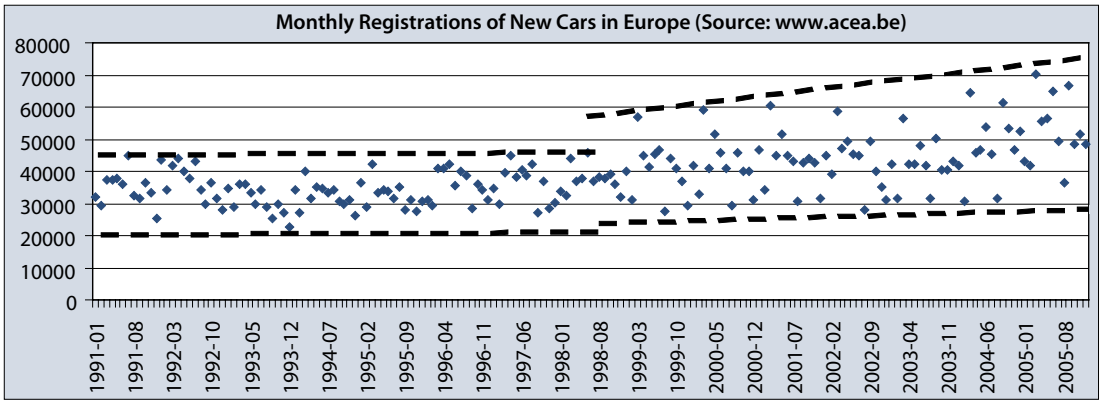


Figure 3. Monthly New Car Registrations in Europe 1991-2005 (Source ACEA)



downward, and no clear trends. Once again, there was no indication of a consistent increase or decrease in the uncertainty of the forecasting environment.

A somewhat different picture emerges when we look at the data (Figure 3) on new car *registrations*. The dashed lines show upper and lower boundaries. Here there seems to be an increase in volatility beginning in 1998. The data shown are officially recorded, monthly registrations of new vehicles of the company in the case study (source: http://www.acea.be).

So the historical data present a mixed picture, with only the car registration time series revealing a pattern of increasing volatility.

CONCLUSIONS

During the 15-year period 1991-2005, the automotive company was able to improve its forecasts for a few countries, a few models, and a few time periods. However, the overall record does not support a trend toward improved forecast accuracy. Rather the results suggest that forecast-accuracy improvements were transient and vanished as the markets changed. Perhaps, the automobile company should have given more attention to its markets, investing in flexibility to react and adapt quickly.

In this paper, I have offered an analytical framework that can be applied to your own company to depict its forecasting track record over time. If you understand your past forecasting performance, you'll be better prepared to face future challenges in setting and achieving forecast-accuracy goals.

REFERENCES

Armstrong, S. (2006). Findings from evidence-based forecasting: Methods for reducing forecast errors, *International Journal of Forecasting*, 22, 583-598.

Byrne, P.J. & Heavey, C. (2006). The impact of information sharing and forecasting in capacitated industrial supply chain: A case study, *International Journal of Production Economics*, 103, 420-437.

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

Küsters, U., McCullough, B. & Bell, M. (2006). Forecasting software: Past, present and future, *International Journal of Forecasting*, 22 (3), 599-615.

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

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

Minnucci, J. (2006). Nano forecasting: Forecasting techniques for short-time intervals, *Foresight: The International Journal of Applied Forecasting*, Issue 4, 6-10.

Sanders, N. & Manrodt, K (2003). Forecasting software in practice: Use, satisfaction, and performance, *Interfaces*, 33 (5), 90-93.

CONTACT
Robert Rieg
Faculty of Business, Aalen University
Robert.Rieg@htw-aalen.de

COMMENTARY ON BENCHMARKING

Teresa McCarthy, Donna Davis, Susan Golicic, and John Mentzer

Robert Rieg and Stephan Kolassa have described what they believe are shortcomings of surveys of forecast accuracy. Each makes reference to our own longitudinal study (McCarthy et al., 2006), and so we welcome this opportunity to reply, as well as to address the broader question of the wisdom of benchmarking forecast accuracy.

Our study explored how sales forecasting management and practice have changed over the past two decades. We replicated and compared results of a survey that was administered to forecasting executives 10 and 20 years prior, while including additional questions to capture new information relevant to the changing business environment. We hoped to provide forecasting managers with a comprehensive view of current and past forecasting practices, to help them understand forecasting trends, and to improve forecasting performance in their own firms.

Our survey explored four overarching dimensions of the forecasting process: forecasting management, techniques, systems, and performance measurement, the last section including data on forecast accuracy. Among the many results presented, our survey revealed that forecast accuracy appears to be deteriorating over time.

SURVEY VALIDITY

Kolassa and Rieg both question the validity of our study's results as benchmarks, partly because we used a survey to collect our data. We agree that surveys have limitations. However, all research methods have their weaknesses. Each also has strengths, and choosing a method to collect and analyze data on any topic always involves a trade-off. McGrath noted years ago that the research process should be regarded "not as a set of problems to be solved, but rather as a set of dilemmas to be lived with," and there is "no one true method that will guarantee success" (1981, p. 179).

The recommended approach is to match the research objective with the most appropriate research method so that the strengths can be maximized and weaknesses minimized. Our research priority was to examine general practices over time in various areas of forecasting management. Therefore, we felt that

- Teresa M. McCarthy, left, is an Associate Professor of Marketing at Bryant University in Smithfield, RI, USA.
- Donna F. Davis is an Assistant Professor of Marketing at Texas Tech University in Lubbock, TX, USA.
- Susan L. Golicic is an Assistant Professor of Supply Chain Management at Colorado State University in Fort Collins, CO, USA.
- John T. Mentzer is the Bruce Chair of Excellence Professor of Marketing and Logistics at the University of Tennessee in Knoxville, TN, USA.

conducting a survey of multiple forecasters from different companies and industries was the best way to obtain this information.

One specific weakness of survey research is the potential for *key informant bias*, since it is difficult to control who actually responds to a questionnaire. We tried to minimize this bias by following recommended survey protocols (Stanton & Rogelberg, 2001). We required a respondent password to complete the survey, ensured that the responses came from companies on our sample list of forecasting executives, and verified that the respondents were in positions affording them knowledge of the forecasting process. It is possible that some of the responses we received were not "truly accurate," but we are confident that informant bias is not a significant challenge to our findings.

Kolassa and Rieg each note that the three surveys did not have identical respondents. We never intended to follow the same companies and industries across the decades. Rather we tried to obtain representative data on the practices companies use in forecasting along with changes in these practices over time. In order to compare practices, we replicated the survey questions about those practices across the surveys and added questions pertaining to new practices that were introduced during the 20-year period.

Kolassa points out that the number of respondents decreased from the earlier studies to the 2006 survey. Unfortunately, response rates to business research in general are declining, due to constraints on practitioners' time and the frequency of requests for their participation in forecasting research. However, lower response rates are acceptable, provided rigorous methods are followed and a satisfactory level of data is obtained (both true of our survey).

Rieg expressed concern about selection bias; that is, any sample of current companies and managers will naturally contain more successes than failures. However, concentration on a single company does not avoid this bias per se, particularly if it is a successful company. Denrell (2005) points out that reducing this bias means working with firms that have failed or are in emerging industries as opposed to a mature industry. Our survey questions sought to provide an accurate picture of current forecasting management, whether the practices were considered sophisticated or dysfunctional. Indeed, our study finds that many aspects of forecasting management have not improved. Instead, we believe they reveal an unsettling downward trend, in spite of increased investments and improved technologies.

BENCHMARKING FORECAST ACCURACY

Both Kolassa and Rieg question the usefulness of benchmarking forecast accuracy as a way to improve forecasting management. We share this concern. Our study does not recommend that reported forecast accuracy results be used as benchmarks.

Direct comparisons of forecast-accuracy levels across firms and industries suffer from several problems. Kolassa raises a key question: "Could we, in principle, collect more or better data and create 'real' benchmarks in forecasting?" We concur that such efforts would be misguided. There is no magic number that qualifies as the correct target for forecast accuracy across organizations, product types, time horizons and/or granularity.

However, we think that collecting and publishing reports of forecast accuracy is nevertheless useful to build knowledge about linkages between forecast accuracy and forecast management. It is also useful to have some indication of increasing or decreasing levels of forecast accuracy for units of analysis beyond a single strategic business unit, such as corporate and industry level analyses.

Forecast-accuracy measurements are key performance indicators for evaluating a firm's forecasting

competence. Managers are obliged to set expectations about appropriate forecast-accuracy goals and to measure progress toward those goals. Ultimately, determining the right forecast-accuracy target is an essential link in aligning business processes with the firm's business needs.

While it is advisable for managers to consider their firm's particular business requirements in setting forecast-accuracy targets, it is important to recognize that business competition is a comparative phenomenon. That is, performance is not judged in isolation. To assure survival, a firm must perform better than competitors. Thus managers want to answer the question, "How do we stack up against the competition?" While we agree that reliance on published reports of forecast accuracy is not appropriate for setting an individual firm's forecast-accuracy targets, we believe that such reports may help managers determine if their targets are viable.

Kolassa argues that reports of forecast accuracy across industries are not helpful, due to noncomparability of industry factors. Yet benchmarking research suggests that companies should look outside their own industries to find best practices that can be adapted to help them gain a competitive edge (Zairi & Al-Mashari, 2005). Innovative approaches to managing processes and people often emerge from an external focus structured to identify, transfer, and adapt best practices in industries other than one's own. The aim of benchmarking research is not to set benchmarks for individual firms but to provide a source of data that, when considered in combination with other sources, can inform process improvement efforts. As noted by Kolassa, "failure to provide benchmarks does not mean the results are uninformative to practicing forecasters" (pp. 9-10).

CONCLUSIONS

Ultimately, the conclusions of our research do not differ substantially from the conclusions made by Kolassa and Rieg. For example, Kolassa writes, "Forecast accuracy improvements due to process and organizational changes should be monitored over time" (p. 14). The implication is that forecast accuracy is just one of many elements to consider and monitor when managing the forecasting process. Similarly, Rieg's general premise is that reliance on improved forecasting algorithms, hardware, and software alone without attention to managing the people, processes, and changes in the external environment could restrict improvements in forecast accuracy. Our research supports both of these conclusions.

Business executives and forecasting managers frequently ask, "What should our forecast accuracy be?" The answer: it depends. Decisions on targeted forecast-accuracy levels must consider multiple factors, such as expected customer service levels, the competitive environment, the resources available within the firm, and existing forecast accuracy in the firm (i.e., a continuous process improvement approach). But no single piece of research on its own can understand and explain all of the intricacies of forecasting management. Our survey is only one piece of this puzzle.

REFERENCES

Denrell, J. (2005). Selection bias and the perils of benchmarking, *Harvard Business Review*, 83 (4), 114-119.

McCarthy, T.M., Davis, D.F., Golicic, S.L. & Mentzer, J.T. (2006). The evolution of sales forecasting management: A 20-year longitudinal study of forecasting practices, *Journal of Forecasting*, 25 (5), 303-324.

McGrath, J.E. (1981). Dilemmatics: The study of research choices and dilemmas, *American Behavioral Scientist*, 25 (2), 179-210.

Stanton, J.M. & Rogelberg, S.G. (2001). Using Internet/Intranet web pages to collect organizational research data, *Organizational Research Methods*, 4 (3), 200-217.

Zairi, M.& Al-Mashari, M. (2005). The role of benchmarking in best practice: Management and knowledge sharing, *The Journal of Computer Information Systems*, 45 (4), 14-31.

CONTACT
Teresa McCarthy
Bryant University
tmccart4@bryant.edu

COMMENTARY ON BENCHMARKING

Jim Hoover, *Foresight* Software Editor

Benchmarking is a concern of critical importance to forecasters and their organizations. The Kolassa and Rieg articles in this issue pose the key questions: Are your forecasts as good as they could be? Is forecast accuracy improving or diminishing over time? And how do you seek relevant information on these questions?

Stephan Kolassa's article gives some needed perspective on the usefulness of published surveys as benchmarks. One problem he identifies is that not all companies measure forecast accuracy using the same metric, but doing so is critical for comparability. Another issue is the low response rate behind the forecasting benchmark surveys. Most broadly, he discusses how differences across product lines, aggregation levels, and time frames for the forecasts all can undermine reliability of the supposed benchmarks.

Adding up the problems in making apples-to-apples comparisons, Stephan concludes that you should not use *external* benchmarks to determine if your own forecasting process is effective. So how *do* you judge its effectiveness? Robert Rieg suggests *internal* benchmarks – measuring changes in a company's own forecast accuracy over time. The internal focus addresses many of the comparability issues. If you are making genuine improvements in forecasting, the results – when measured against prior periods for the same type of item – should show it.

Yet, as Robert argues, external events may cause stable or improving forecast accuracy to deteriorate. In the U.S. Department of Defense, we saw demand for previously stable items rise significantly during the work-up period for the war in Iraq and then decline sharply after the initial ground campaign had ended. Most forecasting methods will have difficulty reacting quickly and appropriately to such external factors.

Robert presents a case study describing a series of forecast cycles in the European automobile industry and the resulting forecast-accuracy outcomes. For this auto manufacturer, after an initial period of improvement, forecast accuracy leveled out and then worsened. Even so, Robert's article illustrates how to make measurements of your own forecast accuracy over time and use them to evaluate and perhaps drive process improvement.

My review of the literature indicates there is very little written on this subject. There are articles on how to measure forecast accuracy, there are principles offered for improving forecast accuracy, and many consultants will tell you they know how to improve forecasting performance. Methodologists tend to show how a new forecasting process fares against other systems when applied to historical data. Why don't they study whether the implementation of a new method improved an organization's forecasting performance over time?

Robert cites a survey by Sanders and Manrodt (2003), which concludes that "the use of appropriate software can lay the groundwork for improved forecasts." I agree. Software can help forecasters avoid some mistakes, such as entering incorrect data, failing to

Jim Hoover is a Captain in the U.S. Navy Supply Corps. He has over 24 years of experience in Supply Chain Management and currently is Chief of Staff at the Naval Supply Systems Headquarters. Previously, as Deputy Commander, Fleet Logistics Operations, he was responsible for world-wide implementation and operation of the Navy's food, parts, retail, fuel, and ammunition supply chains. Jim's software columns and articles appear in Issues 1, 2, 4, and 7 of *Foresight*.

check for outliers, ignoring hierarchical relationships, and the like. So software can potentially improve forecasts. But without measuring your own results over time, how would you know?

I recently worked for an agency with more than $30 billion in annual sales. For years, this organization didn't track forecast accuracy in a systematic way. As part of an expensive implementation of Enterprise Resource Planning (ERP) and a Supply Chain Management System, it decided that tracking forecast accuracy was essential in order to achieve the inventory savings expected from the new system.

Many of the issues Stephan and Robert describe have been experienced by my organization: which specific metrics to use, which forecasting methods to use, how to aggregate individual SKUs' forecasts, how to track the results over time, where to store the forecasts and the actual demands to best allow forecasters access to the results and to propose improvements, and, finally, how to prioritize which of the SKUs they should pursue for accuracy improvements.

My organization had the necessary leadership backbone to attack these problems and still found some seemingly intractable. Despite having a contracted ERP integrator help with the task, it is just now beginning to institute a systematic approach to solving these issues. Progress began when the IT department, operations, and customer relations were brought together to resolve these issues collectively. Committed leadership, oversight, and participation were critical to finally making headway.

CONTACT
Jim Hoover
Naval Supply Systems Command
Hooverjh@aol.com

UPCOMING

Special Feature on Forecastability
Some items are more difficult to forecast than others, so that a particular average percent error may represent a successful outcome for some items and a failure of forecasting for others. The forecastability articles will address these issues and more:

- How can we determine the forecastability of an individual time series, so that we have a basis for judging the success of any forecasting method?
- Relative error metrics have been offered that compare the forecast errors of a designated method to those of a naïve method. The no-change (naïve 1) forecast is a virtual standard in business-forecasting software. Should it be? Are there better alternatives?
- Coefficients of variation on detrended, deseasonalized data are being considered by some companies as indicators of forecastability. Is this metric useful?

Forecast Process Improvement
- The Forecasting Mantra
- Sales Forecasting: Improving the Relationship Between Demand People and Supply People

Forecast Accuracy Measurement:
- How to Define the MAPE
- Measuring Forecast-Improvement Initiatives

The World of Forecasting
Predicting Recessions

Forecast Model-Building
Statistical Significance Testing: Is It Useful?

- Software Reviews
- Book Reviews
- Hot New Research Column
- Forecasting Intelligence Column
- Case Studies

Dear *Trends* Reader,

I'm glad you took the time to download our Special Feature on *Benchmarking Forecasting Accuracy*. You'll find there valuable insights about the dangers of current benchmarking practices and good leads to improve this process.

Each quarterly issue of *Foresight* provides business forecasters with practical, field-tested forecasting advice from top practitioners and world-famous forecasting researchers. I invite you to subscribe to *Foresight* so you, too, can benefit from its concise, objective, and readable articles on important forecasting issues.

Upcoming features include:

- FRAM: The Forecasting Reliability Assurance Model
- A Baker's Dozen Free Sources of Forecasts
- Using Forecasting to Steer the Business: Six Principles
- Why Corporate Culture Matters in S&OP
- Assessing Uncertainty in New Product Forecasts
- The Limitations of Quantitative Modeling

In addition to the magazine, *Foresight* subscribers receive:

- Free online access to our continuously updated library of back issues
- Deep discounts on other *Foresight* publications, including our newest *Anthology,* "How to Evaluate, Manage, and Improve your Forecasting Process" (http://www.forecasters.org/pdfs/foresight/AnthologyContents.pdf)
- Complimentary subscription to *The Oracle*, our forecasting e-newsletter
- Free participation in our forecasting discussion group for Q&A
- Special subscriber discounts to IIF-sponsored forecasting events

To subscribe now, visit http://www.forecasters.org/foresight/subscribe.html.

Very cordially yours,

Len Tashman
Foresight Editor

P.S. Please forward the article to colleagues you think would find it informative, compliments of *Trends* and *Foresight*.