FORESIGHT

THE INTERNATIONAL JOURNAL OF APPLIED FORECASTING

ARTICLES *by* FORECASTERS *for* FORECASTERS

Join the *Foresight* readership by becoming a member of the International Institute of Forecasters
forecasters.org/foresight/

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.