MEASURING FORECASTER PERFORMANCE IN A COLLABORATIVE SETTING WITH FIELD SALES, CUSTOMERS OR SUPPLIER PARTNERS

Hans Levenbach, Ph.D.

Preview:
Measuring performance of forecasters is a complex task, especially with field sales forecasters, customers or collaborative partners as stakeholders in the final forecast. Hans argues that if couched within a systematic forecasting framework, forecasting organizations can achieve greater benefits of accuracy and accountability; hence, gain credibility with management and in their overall approach to forecasting. His recommendation is to use an unbiased, reproducible baseline forecast as an anchor for establishing whether sales persons' forecasts are (1) good, (2) to be reviewed or (3) not acceptable. He describes an easy-to-follow spreadsheet implementation of accuracy measurement for sales forecasters and collaborative partners.

Hans Levenbach is Founder/President of Delphus, Inc., which provides web-based demand forecasting and replenishment planning software solutions for manufacturers, retailers and hospital management organizations. He has extensive experience consulting, training and developing forecasting software applications across multiple industries. Prior to founding Delphus, he was a forecast practitioner and Division Manager at AT&T and research statistician at AT&T Bell Labs. For a couple of decades, Hans has taught forecasting and business statistics courses at Columbia University and New York University. He has served as Board member, President and Treasurer of the International Institute of Forecasters and is an elected Fellow of the IIF. He is the co-author of a textbook *Forecasting – Process and Practice for Demand Management*, published by Duxbury Press in June 2005.

INTRODUCTION

Measuring the performance of forecasters has always been a politically sensitive and complex analytical issue for many firms. There is a familiar adage in corporate organizations that says: "What gets measured gets rewarded, and what gets rewarded gets done." Not only are there human factors involved among organizations ('silos of non-cooperation'), but the mere task of determining an appropriate analytical approach can be forbidding. Nevertheless, in any effective forecasting organization, management needs to measure how well forecasters are doing. This paper describes a new approach that can be used to evaluate forecasts from field sales reps or trading partners in a collaborative environment in which the objective is to arrive at a 'one-number' forecast.

COLLABORATIVE FORECASTING AS A STRUCTURED PROCESS

Traditional forecasting techniques are in widespread use in most companies nowadays. While the software tool-of-choice may be still the ubiquitous MS Excel spreadsheet, forecasters have generally not adopted forecast accuracy measurement and performance reporting as an integral function of the forecasting process. In this paper we introduce a spreadsheet-based tool for reporting forecaster performance in an environment in which sales forecasters are closely collaborating on the forecast with a central organization. In a similar fashion one can evaluate the forecasting performance of trading partners who are closely allied to a firm in providing its inventory forecasts.

What do we mean by collaborative forecasting? In many companies that process is rarely present because so many tend to operate with silos of non-cooperation rather than as a team (Oliva and Watson, 2006). In a collaborative environment, a firm uses a periodic *forecasting cycle* (usually on a monthly basis, but more frequently is also becoming common) to prepare a forecast. This cycle can be described through a number of systematic steps, that we call the **PEER** process: the acronym stands for **P**repare, **E**xecute, **E**valuate and **R**econcile. The **PEER** process is described in detail in Levenbach and Cleary (2005). In this paper dealing with collaborative forecasting we will focus on the **E**valuate and **R**econcile steps.

In the first (Prepare) stage of the process, data is prepared in a suitable relational database containing historical demand, previous forecasts, product, pricing and customer information. After the most recent actuals are posted in the database, central staff produces (Execute) a baseline forecast **BF** (see, for example, Ireland (2005)) over a prescribed forecasting horizon (or lead-time), and for each of hundreds to tens of thousands of planning items or stock-keeping units (SKUs), both in units and revenues. The **BF** is often a statistical forecast or at times a very naïve one comprised of last period's or 'same period previous year' actual demand. Such detailed unit forecasts are required (Evaluate) for production, and aggregate revenue forecasts are utilized by sales, marketing and finance departments. These operational silos - Production, Operations, Marketing, Finance et al. - are often at odds with each other, which dictate that sound Sales and Operations planning (Reconcile) be put in place in order to reconcile diverse forecasts.

Every forecasting organization has its own internal procedures, but it is vital in a collaborative forecasting environment that management proceeds in a structured manner.

There needs to be a sequence of activities that is followed conscientiously by the forecasting staff. If a key step is omitted, either deliberate or inadvertently, his/her credibility can be jeopardized, and credibility is a forecaster's livelihood.

EVALUATING THE PERFORMANCE OF THE **BF** and **SF** FORECASTS

How can we measure forecaster performance in an accountable and equitable manner for salespersons who may have quite different forecasting responsibilities? We begin by analyzing the accuracy of the Baseline Forecast **BF** and the Sales Forecast **SF**. The **SF** should not be confused with the Sales Plan, which is used to set goals for salespeople.

While there are a myriad of metrics one can use (definitely use more than one, preferably ones that can point to different 'downstream' implications), we will focus on the *Absolute Percentage Error* (APE) for baseline forecast **BF** and *Accuracy%* for sales forecast **SF**. If we denote **A** = actual and **BF** = baseline forecast, the APE is calculated from a forecast error with the formula:

$$\text{APE} = 100\% \ (|\mathbf{A} - \mathbf{BF}|/ \mathbf{A}),$$

where the vertical bars denote the absolute value: (|**A** – **BF**| = Absolute value (**A** – **BF**). An APE equal to 0.08, for example, indicates that the **BF** missed the actual demand by 8%.

The accuracy of a sales forecast **SF** is typically measured relative to the forecast, not the actual (See, for example Crum (2003. p. 167). The basis of this measurement is called *Accuracy%* and it can be related to the percentage error PE:

$$\begin{aligned}\text{Accuracy\%} &= 100\% \ [1 - (|\mathbf{A} - \mathbf{SF}|/\mathbf{SF})] \\ &= 100\% \ / \ [(\mathbf{OF} + \mathbf{A})/\mathbf{A} \ - \ (\mathbf{A} - \mathbf{BF})/\text{A} \ ].\end{aligned}$$

The denominator in the formula is **SF** and **OF** represents the override made to the **BF** by the salesperson**:** OF = SF - BF.

Accuracy% is conventionally used for sales forecasts, because the error is measured in reference to the Forecast, not the Actual. In the case of an APE calculation, the rationale is that the error is measured in reference to the Actual.

There is a mathematical relationship between *Accuracy%* and the APE. We will use a summary of the APEs, over a suitable forecast horizon, as an *anchor* to determine whether *Accuracy%* is (1) good, (2) to be reviewed or (3) not acceptable. The anchor is based on the baseline forecast, which we have established by objective means (namely having a basis in a statistical model) to be unbiased, reproducible, and credible to management. Note that when *Accuracy%* yields unrealistic numbers, the software implementation needs to take account of this.

How many forecasts do you need to make: that is, over how many periods must the APE and *Accuracy%* be calculated? Our recommendation is to evaluate over a rolling planning cycle, which is typically a 12 or 18-month period. For new products, there may not be sufficient data to get a reasonable interpretation of forecast accuracy, so less refined measurements should be made.

The type of average used for a baseline anchor is also a consideration. To measure forecast performance we would need the distribution of forecast errors. This is rarely known. Instead, we usually make an assumption of normality which means we only need a mean and a variance to describe the entire distribution. Normality is also rare in forecast errors, except when used in theoretical modeling assumptions. In practice, we prefer to use the median of the APE's – the MdAPE – to the mean of the APE's (MAPE) because it is less sensitive to the occurrence of unusual or extreme values in the forecast errors, the likely skewness of the distribution of forecast errors, and the relatively small number of forecast errors in a calculation (usually fewer than 30).

As an illustration, suppose we wish to evaluate the accuracy of the March 2008 baseline forecast. Over the previous 18 months, starting in November 2006, the forecasting organization will have posted up to 18 baseline forecasts for March 2008; one each month for the month of March 2008. Once March 2008 arrives, we can calculate an Accuracy% for **SF** as well as an MdAPE for **BF**. Then we substitute some percentiles of the empirical APE distribution in the Accuracy% formula to determine what the shade or color needs to be in the report. For **SF**, some weighted average of Accuracy% estimates for a given period can be made instead of using the latest Accuracy%. In any case, the treatment of these estimates should be communicated beforehand and applied uniformly and consistently across the entire sales force. In practice, you would have even fewer **SF** forecasts since sales forecasts are rarely updated monthly over a complete planning cycle. More typically, such an update may occur only semi-annually. Management will need to decide beforehand when to 'lock in' the **SF** forecasts from the salespersons to make the *Accuracy%* calculation. The **BF** should be evaluated using the same 'lock-in' date.

Next, we will describe how our procedure determines a color-coded benchmark value for *Accuracy%* which is then compared to a current *Accuracy%* calculation for the sales forecast for a particular month or summary period in an FS forecast. On the spreadsheet implementation, we have assigned the colors green, amber and red to each grading, respectively. In a report this may show as a light, medium and dark shade, but the three colors are reminiscent of driving behavior in traffic.

OPERATIONAL STEPS FOR MEASURING FIELD SALES PERFORMANCE

Consider a forecaster named Aaron who is responsible for the Sales Forecast for a product family QQQ in his Territory. In the waterfall chart in Figure 1, let us assume that the top line of the first column is the actual for March 2007 (= 1,001,666) for Aaron's product line QQQ. It appears that the **BF** forecasts were almost all over-forecasts (negative sign) and that the MdAPE is 2.2 [= (2.12 + 2.22)/2 = 2.17], not really different very from the MAPE (= 2.1) in this case.

Figure 1. Waterfall Chart Based on Holdout Sample (12 Months) for Product Line QQQ

Waterfall Chart with Holdout Sample (12 months) for Item QQQ

	PE (%)												
Hold-Out	1	2	3	4	5	6	7	8	9	10	11	12	MAPE (%)
1,001,666	−5.38	−2.22	−1.28	−0.70	−2.33	2.53	−0.36	−2.60	−1.72	−0.06	−2.12	−4.07	2.1
1,073,196	−3.85	−1.95	−1.08	−2.55	1.86	0.41	−2.72	−2.52	−0.59	−2.14	−4.73		2.2
1,421,423	−3.58	−1.75	−2.94	1.65	−0.27	−1.92	−2.62	−1.36	−2.66	−4.75			2.4
1,577,321	−3.37	−3.62	1.28	−0.49	−2.64	−1.83	−1.48	−3.46	−5.29				2.6
1,600,991	−5.27	0.62	−0.88	−2.85	−2.54	−0.70	−3.58	−6.12					2.8
1,594,481	−0.96	−1.55	−3.24	−2.75	−1.39	−2.77	−6.23						2.7
1,510,052	−3.16	−3.93	−3.16	−1.61	−3.49	−5.41							3.5
1,436,164	−5.58	−3.84	−2.00	−3.71	−6.14								4.3
1,404,978	−5.49	−2.68	−4.11	−6.36									4.7
1,585,409	−4.32	−4.80	−6.78										5.3
1,234,848	−6.46	−7.49											7.0
923,115	−9.20												9.2

The next step is to segment the items into color zones according to the baseline accuracy metric. For example, if the MdAPE is between 0 – 5%, we assign that to the GREEN zone. The AMBER zone corresponds to those items whose MdAPE is greater than 5% but less than 30% in absolute value. In the RED zone are all those items whose MdAPE is greater than 30%. You can assign these % cutoffs according to criteria suitable to your particular environment. For instance, these cutoffs can be determined according to some 80-20 rule or the percentiles of an empirical distribution.

As we have seen above, we can express *Accuracy%* as a function of two interpretable quantities:

$$\textit{Accuracy\%} = 100\% \,/\, |[\ (\mathbf{OF} + \mathbf{A})/\mathbf{A}\ - (\mathbf{A} - \mathbf{BF})/\mathbf{A}\ ]|$$

Firstly, the quantity (**A** – **BF**)/**A** is the Percentage Error of the baseline forecast **BF** and, secondly, the quantity (**OF** + **A**)/**A** can be interpreted as the *field sales influence* or the *degree of demand shaping* on the actual. If we substitute the MdAPE as our *anchor* for (**A** – **BF**)/**A** in the formula and the **OF** made by Aaron for that month, we obtain a benchmark measure of his *Accuracy%*. His actual *Accuracy%* for that same period can be calculated and compared to this benchmark. For example, if we use Aaron's December 2007 estimate for the March 2008 Sales Forecast forecast and assume, for simplicity, that **OF** = 0, the benchmark *Accuracy%* is 100% / | 1 +/- 0.022| = 102.2 % or 97.8%, depending on the sign of **A** - **BF**. Select Accuracy% benchmark value that is less than 100%. When **OF** = 0, we treat the baseline forecast as the sales forecast. In other words, no overrides or adjustments to the baseline forecast were made by the salesperson.

Similar calculations can be made if **OF** is not zero. In several companies across different industries, we found at least 40% of the SKU-level forecasts had received overrides from sales people. Most of these overrides were made at a summary level, such as a product line, but ended up prorated to the lower SKU-levels for production purposes. The accuracy calculations, however were made at the product line for the sales people.

As another example, Aaron's manager Natasha is considering two brands in Figure 2 that have different volumes and variability of forecast error. It could be easier to forecast Brand Y than Brand_X when you consider the relative variability of forecast errors in each Brand. In this example, it turns out that for Brand X, an *Accuracy%* of 86% places it in the green zone, while that same *Accuracy%* would place it in the Amber zone for Brand_Y. Likewise, an *Accuracy%* of 65% for Brand_Y places one in the Red zone while the same percentage would place one in the Amber zone for Brand_X. This is because of the difference between the relative variability of the baseline forecast errors in Brand_X and Brand_Y. Because Brand Y is less variable, it should be harder to achieve the same *color* rating as Brand X for the same *Accuracy%*.

Figure 2. Comparison of Brand Performance (Product). The sales forecaster's 'lock-in' period is one month.

Brand_X							
Forecast	292,000	292,000	251,858	414,492	349,577	399,500	382,433
Actual	250,870	338,129	415,860	439,100	476,790	424,640	440,560
Accuracy%	86%	84%	35%	94%	64%	94%	85%

Brand_Y							
Forecast	6,949,526	6,418,084	6,740,410	6,609,145	5,377,230	6,001,169	6,437,869
Actual	6,694,447	5,713,863	6,597,193	6,396,258	7,264,969	7,366,172	6,405,431
Accuracy%	96%	89%	98%	97%	65%	77%	99%

Another kind of comparison can be made for different locations like Plants. In this case, salesperson Jordan can be evaluated for his forecast performance in the two Plants that are his responsibility to forecast. As shown in Figure 3. these results show that, eventhough the Plants are of comparable size, it takes a higher Accuracy% to achieve green in Plant A than in Plant B. If you compare May 07 performance in the two Plants, the greater variability in Plant B makes it harder to achieve a 76% than Plant A.

Figure 3. Comparison of Plant performance (Customer/Location). The sales forecasters 'lock-in' period is one month.

	C	D	E	F	G	H	I	J	K	L	M
1			APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
2			'07	'07	'07	'07	'07	'07	'07	'07	'07
3	**Plant A**										
4											
5	Product										
6	999 - Total Product										
7		Forecast	490,105	412,205	667,950	662,200	622,815	583,235	686,245	669,714	319,337
8		Actual	435,156	311,324	682,715	473,081	648,137	621,984	744,442	710,124	647,207
9		% Error	-13%	-32%	2%	-40%	4%	6%	8%	6%	51%
10		Accuracy%	89%	76%	98%	71%	96%	93%	92%	94%	0%
11											
12											
13	**Plant B**										
14											
15	Product										
16	999 - Total Product										
17		Forecast	214,720	366,910	222,000	461,000	750,000	864,000	859,030	807,000	479,420
18		Actual	105,167	280,240	295,540	542,579	656,015	736,857	783,440	485,813	432,297
19		% Error	-104%	-31%	25%	15%	-14%	-17%	-10%	-66%	-11%
20		Accuracy%	49%	76%	67%	82%	87%	85%	91%	60%	90%
21											

In some cases, several sales forecasters may be responsible for the same product, product family or brand, but for their own respective sales territories. In Figure 4 we show a comparison among two field sales forecasters Ivan and Daphne. For readability we don't show the full year, but the report can be extended over an entire planning cycle, say 18 months. We show prior year history as a reference from which one can calculate the relative variability of the data, say by the coefficient of variation. The coefficient of variation is one measure that can establish the relative variability of the product family. For Daphne, this is 0.57 and 0.55 calculated from the history (count = 24) and forecast (count = 12) in the report, respectively. The corresponding statistics for Ivan are 0.33 and 0.28. Hence, the variability for Daphne is almost twice that of Ivan. While Ivan appears to have smaller misses in his one-step ahead forecasts, he also enjoys less variability, so it is harder to get a 'green zone' than Daphne whose Territory is more volatile and for whom it is more difficult to obtain an equally high *Accuracy%* as Ivan. In using these metrics, one has to be cautious with their interpretation. When two individuals get the same score, that does not suggest that they have done an equally well job. The color schemes help to differentiate these scores.

Figure 4. Comparison of field sales forecasters for Product Family ABC. The sales forecasters' 'lock-in' period is one month.

	B	C	D	E	F	G	H	I	J	K	L	M
1			**OCT**	**NOV**	**DEC**	**JAN**	**FEB**	**MAR**	**APR**	**MAY**	**JUN**	**JUL**
2			06	06	06	07	07	07	07	07	07	07
3			SEP 06_Forecast	OCT 06_Forecast	NOV 06_Forecast	DEC 06_Forecast	JAN 07_Forecast	FEB 07_Forecast	MAR 07_Forecast	APR 07_Forecast	MAY 07_Forecast	JUN 07_Forecast
4	SALESPERSON											
5	**Ivan**											
6												
7	Product Family ABC											
8		Prior Year	74,723	65,711	106,731	95,128	104,777	104,187	86,791	61,399	76,477	39,115
9		Forecast	63,959	61,367	60,375	84,627	102,208	92,012	63,611	69,120	124,621	109,995
10		Actual	69,926	69,049	109,622	40,957	91,967	143,194	93,221	122,150	129,312	50,045
11		% Error	9%	11%	45%	-107%	-11%	36%	32%	43%	4%	-120%
12		Accuracy%	91%	87%	18%	48%	90%	44%	53%	23%	96%	45%
13												
14												
15	**Daphne**											
16												
17	Product Family ABC											
18		Prior Year	47,882	131,584	68,425	88,769	81,580	95,525	63,101	63,703	97,594	35,422
19		Forecast	47,464	66,402	58,550	56,288	50,544	68,239	43,580	70,016	68,252	103,385
20		Actual	86,500	43,611	73,645	62,567	75,759	63,169	90,165	87,434	262,371	103,088
21		% Error	45%	-52%	20%	10%	33%	-8%	52%	20%	74%	0%
22		Accuracy%	18%	66%	74%	89%	50%	93%	0%	75%	0%	100%
23												

SOME KEY POINTS TO REMEMBER

- Use forecast errors from unbiased, objective models to anchor forecaster performance
- Consider weighting salesperson accuracy based on nearness of the forecasted month.
- Agree to a 'lock-in' period for the forecasts prior to starting the measurement process
- Apply a robust alternative to conventional measures of accuracy to validate reliability in measurement of variability
- Improve forecast accuracy through continuous evaluation of model and forecaster performance
- For accountability of the final forecast, avoid simply combining forecasts
- A combined forecast, if utilized, can be evaluated with the same criteria as the SF.
- Enhance management credibility through a structured forecasting process

REFERENCES

Crum, C. (2003). *Demand Management Best Practices*. Boca Raton, FL: J. Ross Publishing Inc.

Ireland, R. K. and C. Crum (2005). *Supply Chain Collaboration*. Boca Raton, FL: J. Ross Publishing Inc.

Levenbach, H. and J. P. Cleary (2005). Forecasting – *Practice and Process for Demand Management*. Belmont, CA: Duxbury Press.

Oliva, R. and N. Watson (2006). Managing functional biases in organizational forecasts. *Foresight: The International Journal of Applied Forecasting*, Issue 5, 27 – 31.

Contact Info:

Hans Levenbach
Delphus, Inc.
hlevenbach@delphus.com.