MEASURING FIELD SALES PERFORMANCE: A COLLABORATIVE APPROACH TO IMPROVING FORECASTING ACCURACY

By Hans Levenbach

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

Dr. 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 manager at AT&T and research statistician at AT&T Bell Labs. During his spare time, Hans has taught forecasting and business statistics courses at Columbia University and New York University. Hans is an elected Fellow of the International Institute of Forecasters. He is the co-author of a textbook *Forecasting – Process and Practice for Demand Management*, published in June 2005.

Measuring the performance of forecasters has always been a politically sensitive and complex analytical issue for many practitioners. 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 forecaster performance from field sales reps to trading partners who are involved in a collaborative forecasting and planning environment (see Ireland (2005) for a discussion of implementing a collaborative forecasting process) in which the firm's objective is to arrive at a ‘one-number’ forecast in a Sales & Operations (S&OP) process.

COLLABORATIVE FORECASTING AS A STRUCTURED PROCESS

When I was approached by several forecasting organizations in quite different industries about a way to evaluate forecasters (more specifically, the adjustments made to a baseline forecast by field sales forecasters in their territories), it reminded me of past times when, as a forecaster for AT&T, we were interested in evaluating field forecasters from the Bell System Operating Companies. Then, computing was still costly and time consuming. Nowadays, computing is fast and cheap, and statistical forecasting tools are in widespread use in most companies. While the software tool-of-choice in many companies may be still the ubiquitous MS Excel spreadsheet, forecasters have been reluctant to adopt forecast accuracy measurement and performance reporting as an integral function of the forecasting process.

In this paper I introduce a spreadsheet-based tool for reporting forecaster performance in an environment in which sales forecasters (including customers and trading partners) are closely collaborating on the demand forecast with a central organization. In a similar fashion I can evaluate the forecasting performance of customers and trading partners who are closely allied to a firm in providing its inventory forecasts.

In many companies today, the Sales and Operations (S&OP) process has been implemented as a way of collaboratively balancing supply and demand subject to financial objectives. The S&OP initiative, while not new, represents a major opportunity for forecast improvement in that collaboration on forecasting is rarely present. So many organizations tend to operate as silos of non-cooperation rather than as a corporate team (Oliva and Watson, 2006).

In a collaborative environment, a firm uses a periodic *forecasting cycle* (usually on a monthly basis, but more frequently weekly is also becoming common) to prepare a forecast. This forecasting cycle involves a number of systematic steps that I call the PEER process. In the first (**P**repare) 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, a forecasting staff is then able to produce (**E**xecute) a baseline forecast **BF** (see, for example, Ireland and Crum (2005)) over a prescribed forecasting horizon (or lead-time). This is done for each of hundreds to tens of thousands of planning items or stock-keeping units (SKUs), both in units and

revenues, for hundreds of locations. The detailed unit forecasts then are analyzed (**E**valuate) in scorecards and distributed in dashboards or reports for inventory planning and production scheduling. In addition, revenue forecasts at various aggregations are utilized by sales, marketing and finance departments for their planning needs. These organizational silos - Production, Operations, Marketing, and Finance - are often at odds with each other about the forecast to go forward with for the company. This necessitates that a sound Sales and Operations planning (**R**econcile) process be in place in order to reconcile diverse forecasts into a single baseline forecast for use by planners.

The baseline forecast **BF** is often a statistical forecast but sometimes can just be a naïve forecast comprised of last period's or 'same period previous year' actual demand. The latter can happen if units are contracted amounts or intermittent service parts for inventory planning.

PEERING INTO THE FUTURE WHILE COLLABORATING WITH PEERS

If you now take the first letters in the four bold-faced stages in the paragraph above, you will see that it spells **PEER**. One way of looking at this is that you are "peering into the future" while collaborating with your peers. Every forecasting organization may have its own internal procedures, but it is vital in this forecasting environment that forecasters proceed collaboratively 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.

In this paper dealing with collaborative forecasting I will focus on the **E**valuate and **R**econcile steps of the **PEER** process. The overall **PEER** process is described in more detail in Levenbach and Cleary (2005).

EVALUATING FORECASTER PERFORMANCE

How can we measure forecaster performance in an accountable and equitable manner for salespersons who may have quite different forecasting responsibilities and objectives? I begin by using the accuracy measure of the Sales Forecast **SF** and relate it to the variation in the Baseline Forecast **BF.** In practice, the field sales people are asked to make their adjustments to the baseline forecast based on their knowledge of the market and anticipated sales to their customers.

In sales organizations, the "accuracy" of a sales forecast **SF** is typically measured relative to the sales forecast **SF**, not the actual **A** (See, for example Crum (2003. p. 167). The basis for this measurement is called *Accuracy%* and is defined by:

$$\text{Accuracy\%} = 100\% \, [1 - (|\mathbf{SF-A}|/\mathbf{SF})]$$

where the vertical bars denote absolute values.

While there are a myriad of metrics one can use (I recommend you use more than one, if appropriate, and preferably ones that can assess different 'downstream' implications), you will see that the Accuracy% measure is mathematically related to the baseline forecast **BF** occurring in the *Percentage Error* PE. If we denote **A** = actual and **BF** = baseline forecast, the PE is defined by the formula:

$$PE = 100\% \ (\mathbf{A} - \mathbf{BF})/ \mathbf{A},$$

A PE equal to 0.08, for example, indicates that the baseline forecast missed the actual demand by 8%. The PE can be rewritten as

$$PE = 100\% \ / \ [(\mathbf{OF} + \mathbf{A})/\mathbf{A} - (\mathbf{A} - \mathbf{SF})/\mathbf{A}],$$

where **OF** represents the override or adjustment made to the baseline forecast by the salesperson**:** SF = BF + OF.

There is a mathematical relationship between *Accuracy%* and PE. With some algebraic manipulations, the Accuracy% measure is related to the PE by the formula

$$\text{Accuracy\%} = 100\% \ \{1 - \ \big| \ [1/ ((\mathbf{OF} + \mathbf{A})/\mathbf{A} - PE)] -1 \ \big| \ \}$$

I will use a summary of the PEs, over the appropriate forecast horizon, as a *range* to determine whether *Accuracy%* is (1) good, (2) to be reviewed or (3) not acceptable. This range is based on baseline forecast accuracies, which I can establish 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 PE be calculated? My recommendation is to perform this evaluation 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 measurements with analogous products should be made.

To measure forecast accuracy completely we would need know the distribution of forecast errors. This is rarely, if ever, available. Instead, one usually makes the assumption that forecast errors are normally distributed which means that only a mean and a variance are needed to describe the entire distribution. I have found normality to be a rare occurrence in forecast errors, except in theoretical models.

In practice, I therefore prefer to use the median of the absolute PE's (MdAPE) to complement and check the validity of the mean of the absolute PE's (MAPE) as a summary, because the former is much less sensitive to the occurrence of unusual or extreme values in forecast errors. The problem is aggrevated by the likely skewness of the

distribution of forecast errors, and the relatively small number of forecast errors available in a practical situation (usually fewer than 30).

As an illustration, suppose we wish to evaluate the accuracy of a particular period, say the March 2008 baseline forecast. Over the previous 18 months, starting in November 2006, the forecasting organization will have been able to post up to 18 baseline forecasts for March 2008; one each month made for the month of March 2008. Once March 2008 rolls around, 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 used instead of just 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, there are usually even fewer **SF** forecasts since sales forecasts are rarely updated monthly over a complete planning cycle. Most field sales forecasters need to be coerced into participating in a monthly forecast. Their focus is on their sales goals and the customer. More typically, a field sales forecast 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, I 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 a collaborative field sales forecast. On the spreadsheet implementation, I 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

Aaron, a field sales forecaster, is responsible for adjusting a unit baseline 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 very different from the MAPE (= 2.1) in this case. It is good practice to report both the MAPE and the MdAPE. When they are close, as in this case, it is safe to report the MAPE, otherwise report the MdAPE as the preferred measure of a representative absolute PE .

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 measure. For example, if the MdAPE is between 0 – 5%, I 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%. With management participation, you can assign these % cutoffs according to criteria suitable to your particular environment. For instance, the 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:

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

Firstly, the quantity 100%*(**A** – **BF**)/**A** is the Percentage Error PE 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 I substitute the MdAPE as our *anchor* for (**A** – **BF**)/**A** in the formula and the **OF** made by Aaron for that month, I 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 I use Aaron's December 2007 estimate for the March 2008 Sales 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, I 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, I found that 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, even though the Plants are comparable in 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, I show a comparison among two field sales forecasters Ivan and Daphne. For readability I don't show the full year, but the report can be extended over an entire planning cycle, say 18 months. I show prior year history as a reference from which one can calculate the relative variability of the data, say by a coefficient of variation CV (CV = sample standard deviation/sample mean). 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, it does not always mean 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 central tendency and variability
- Improve forecast accuracy through continuous evaluation of model and forecaster performance
- For accountability of the final forecast, avoid simply weighting forecasts equally
- 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. (2005). ABC of Collaborative Planning, Forecasting and Replenishment. The Journal of Business Forecasting (24, issue 2).

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.
URL: www.delphus.com
Email: hlevenbach@delphus.com