Hospitality Review

Volume 17
Issue 1 *Hospitality Review Volume 17/Issue 1*

Article 4

1-1-1999

Monitoring the Accuracy of Multiple Occupancy Forecasts

Zui Schwartz
Ben-Gurion University of the Negev, Israel, null@in.bgu.ac.il

Monitoring the Accuracy of Multiple Occupancy Forecasts

Abstract
Corporate executives closely monitor the accuracy of their hotels' occupancy fore- casts since important decisions are based upon these predictions. This study lists the criteria for selecting an appropriate error measure. It discusses several evaluation methods focusing on statistical significance tests and demonstrates the use of two adequate evaluation methods: Mincer- Zamowitz's efficiency test and Wilcoxon's Non-Parametric Matched-Pairs Signed- Ranks test.

Monitoring the accuracy of multiple occupancy forecasts

by Zvi Schwartz

Corporate executives closely monitor the accuracy of their hotels' occupancy forecasts since important decisions are based upon these predictions. This study lists the criteria for selecting an appropriate error measure. It discusses several evaluation methods focusing on statistical significance tests and demonstrates the use of two adequate evaluation methods: Mincer-Zarnowitz's efficiency test and Wilcoxon's Non-Parametric Matched-Pairs Signed-Ranks test.

Hotels incorporate various forecasts into their control cycle.[1] These predictions differ in the forecasted entity (rooms sold, revenues, profits), their basic unit (a day, a month, or a year), their forecasting horizon (how many periods ahead are forecasted), and their frequency (how often the forecast is updated). While efficient managerial decisions require that all these forecasts be accurate,[2] the accuracy of the occupancy forecasts is of special importance because all the hotel departments, not only rooms, rely upon accurate occupancy forecasts.

Occupancy forecasts serve as building blocks for the forecasts of the hotel's usage of restaurants, telephone, garage, laundry, business center, and audio/visual rental because the performance of these and other departments is highly correlated with the number of rooms rented. Both over-forecasting and under-forecasting can significantly impair the hotel's profitability. Over-forecasting leads to overspending and insufficient marketing efforts. Under-forecasting results in under-staffing, which leads to poor service and increased stress among employees. Consequently, hotel managers consider occupancy forecasts to be a valuable part of their control cycle, and an objective evaluation of the accuracy of these forecasts is of the utmost importance.

The evaluation of forecast accuracy is performed frequently. In well-managed hotel chains, it is done by corporate executives.

FIU Hospitality Review, Volume17, Number 1 & 2, 1999

Every month, the hotels submit their monthly forecast to their corporate office and the next month they generate a follow-up report, often called the variance report,[3] which lists the actual daily occupancy figures against the predictions. Corporate executives use it to closely monitor the accuracy of the occupancy forecasts. Ideally, they should be able to identify hotels which routinely submit inaccurate forecasts and take measures to alleviate the problem.

Evaluation is difficult

It is a challenging task to evaluate a large number of detailed occupancy forecasts in an objective and practical manner. Periodical monitoring involves a large number of hotels; for some companies it means hundreds of accuracy tests per month. It is therefore important to devise an efficient monitoring method, that is, a computerized system that could handle a large number of occupancy forecasts. There are various evaluation models that are appropriate for hotel occupancy forecasts. Their relevance is assessed based on the following general guiding rules:

- The evaluation method must be statistically sound to guarantee the reliability and validity of the results.
- The evaluation method should reflect the characteristics of the investigated phenomena. That is, the error measure should match the hotel's estimated cost of error. A failure to apply an adequate evaluation method might result in a misleading characterization of the assessed forecast.
- The method employed should match the purpose of the evaluation exercise. An evaluation method that is appropriate for in-depth analysis of a single set of occupancy predictions in a single hotel might not be useful to a corporate office conducting its periodical accuracy tests for a large number of hotels. The purpose of the in-depth analysis is to improve the accuracy of a specific forecasting model. The goal of the monthly monitoring routine is to ensure that the forecast accuracy of each of the chain's hotels does not fall below an acceptable level. If it does, the corporate office notifies the hotel, with the expectations that the hotel's managers will improve the accuracy of future forecasts.

Hence, for a corporate monitoring task, an adequate evaluation method must be automated, i.e., programmable, and produce an output that clearly indicates whether the accuracy of each hotel's forecast meets the company's acceptable level.

The evaluation of the forecast's accuracy is most often based on the forecast's error. For a single observation denoted by t, a forecast error, E_t, is the difference between the forecast f_t and the actual value A_t, that is, $E_t = f_t - A_t$. Hence, a t period forecast generates t forecast errors. Various measurements which are based on these errors are used by scholars and practitioners.[4] Most common are the mean absolute deviation, mean squared error, sum of squared error, mean absolute percentage error, and standard deviation of error (See Appendix A for listing of formulae).

Recent forecasting studies in the hospitality industry have applied these traditional error measures to evaluate forecasting models: Miller, McCahon, & Miller[5] adopted the criterion of minimizing Mean Absolute Deviation (MAD) and the Mean Squared Error (MSE); Andrew Cranage and Lee[6] and Cranage and Andrew[7] looked mainly at the Sum of Squared Errors (SSE); Messersmith and Miller[8] recommended using Bias, MAD, and MSE; Schwartz[9] applied the Absolute Percentage Error (APE); and Schwartz and Hiemstra[10] used the Mean Absolute Percentage Error (MAPE).

Cost of error measured

It is important to select an appropriate error measure since the type of error measure might determine the result of the forecast performance evaluation. The error measure should reflect the estimated damage caused by the forecast error, i.e., the cost of error. The popular "squared" error measures (e.g., SE, MSE, SSE) assume that as the forecast error increases, the resulting damage to the hotel increases exponentially. This is the same as saying that it is worth more to reduce the forecast error by, for example, 10 rooms when the error is 50 rooms, than to reduce the error by the same amount (10 rooms) when the error is 20 rooms. It is not clear whether this is indeed the cost structure in all hotels. A typical hotel production cost function, which serves as a reasonable indicator for the cost structure of an over-forecast error, is often more linear than exponential. The cost of under-forecast has more to do with dissatisfied guests and overworked employees, and as such is more difficult to generalize.

In some cases, the cost of the forecast error is asymmetrical; that is, the cost of an under-forecast error might be larger (or smaller) than the cost of an over-forecast error. Asymmetry is often the result of differing cost functions; the cost function of the under forecast error is not identical to the cost function of the over forecast (e.g., linear vs. exponential). If present, this asymmetry must be properly addressed by the error measure. The following equation can serve as the lost function when the cost of under-forecast is exponential and the cost of over-forecast is linear:

$$C(E) = \phi\left(E_t - |E_t|\right)^2 + \varphi\left(E_t + |E_t|\right)$$

where the parameters ϕ and φ (ϕ, $\varphi > 0$) require an estimation.

Ratio (percentage) based error measures such as APE and MAPE are useful when comparing forecasts of different scales. This is especially important if the corporate office wishes to compare the forecast accuracy across hotels with varying capacities.

Once an appropriate error measure has been determined, an evaluation method must be selected. There are two types of evaluation methods: absolute and comparative.

Methods assess absolute value

Several methods attempt to assess the forecast's "absolute" value. Among them are Theil's decomposition, Kolmogorov's "optimal" forecast, and Mincer-Zarnowitz's efficiency.

- **Theil's Decomposition:** Theil[11] decomposed the sample's average squared forecast error:

$$D_N^2 = (\bar{f} - \bar{A})^2 + (S_f - rS_s)^2 + (1 - r^2)S_x^2,$$

where $\bar{f}$ denotes the sample mean of the forecast, $\bar{A}$ denotes the sample mean of the actual series, S_f denotes the sample standard deviation of the forecast, S_s denotes the sample standard deviation of the actual series and r denotes the sample correlation between forecast and the actual series. Theil proposed the following three proportions:

$$U^M = \frac{(\bar{f} - \bar{A})^2}{D_N^2}$$

, where U^M compares the means of the observed and predicted series.

$$U^R = \frac{(S_f - rS_A)^2}{D_N^2}$$

, where U^R measures the extent to which the slope of a regression (actual as a function of the predicted values) is different from unity.

$$U^D = \frac{(1 - r^2)S_A^2}{D_N^2}$$

, where U^D assesses the relative size of the regression's error term.

For optimal forecast, the proportions U^M and U^R should be close to zero and U^D should be close to unity.

The major drawback of Theil's method is its subjective nature: It does not provide an objective statistical significance test. One does not know how small the proportions U^M and U^R should be to indicate an accurate occupancy forecast. While one can arbitrarily decide on an acceptable level for each of the three measures, there is no statistical measure to assess the significance of the results.

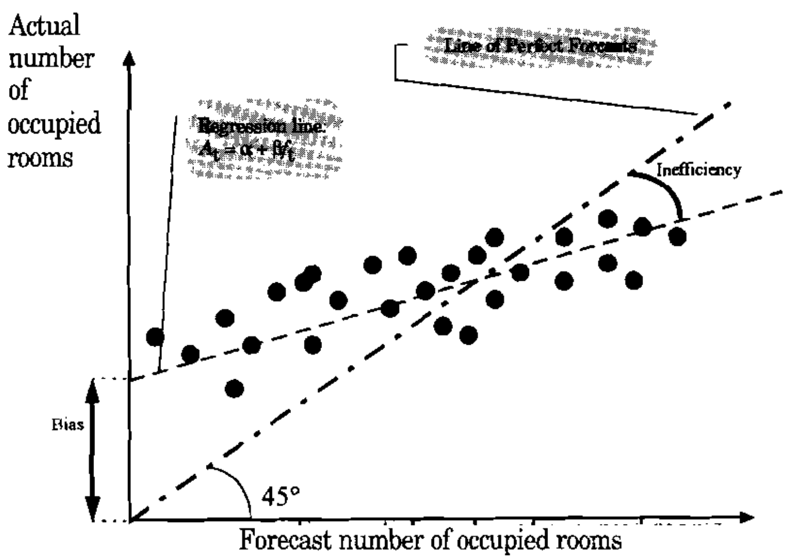
- **The "optimal" forecast:** Janacek[12] shows that given a set of data (actual), one can estimate the performance of an "optimal" forecast using the Minimum Attainable Forecast Error Variance (MAFEV). The MAFEV[13] is given by:

$$\sigma^2 = \left[(2\pi)^{-1} \int_{-\pi}^{\pi} log\{2\pi f(\omega)\}d(\omega)\right],$$

where $f(\omega)$ denotes the power spectrum.

- **The Mincer-Zarnowitz Efficiency Test:** Mincer and Zarnowitz[14] regress the actual observations (denoted by A_t) against the predicted values (denoted by f_t): $A_t = \alpha + \beta f_t + \varepsilon_t$, where t=1,2,...,n, and ε_t is the regression's error term. When the forecast is perfect, the plotted regression line is identical to the 45° line in Figure 1. That is, α=0,

Figure 1
Line of perfect forecasts and the regression line of actual vs. Forecast



β=1 and E(ε) = 0. α≠0 indicates that the forecast is biased and a ((1 indicates that the forecast is inefficient. The joint null hypothesis that a forecast is efficient and unbiased is H_0: α=0, β=1.

F_1 the test statistic for the bias/efficiency test verifies that the bias and the inefficiency unveiled by the chart is statistically significant:

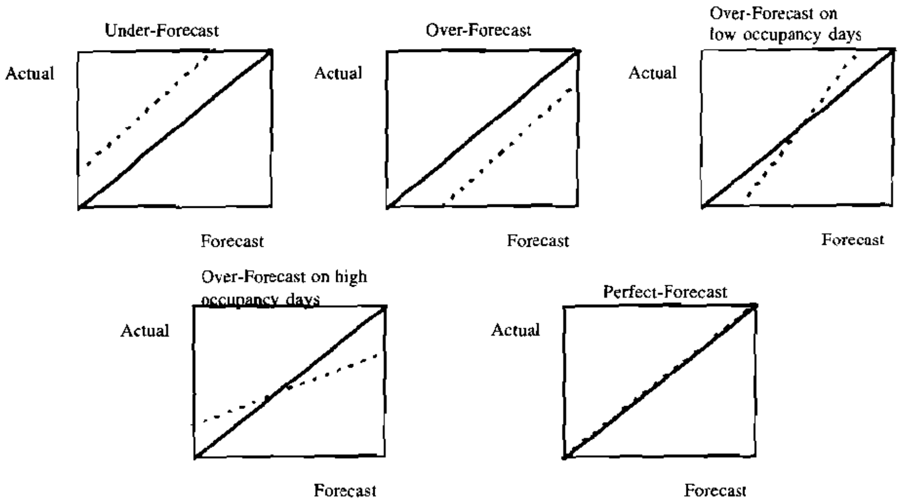
$$F = \left[\sum_{t=1}^{n}(A_t - f_t)^2 - \sum_{t=1}^{n}(A_t - a - bf_t)^2 \frac{1}{2}\right] \Big/ \left[\sum_{t=1}^{n}(A_t - a - bf_t)^2 \frac{1}{n-2}\right],$$

where n is the number of observations, and a and b are the least square estimates of α and β. The null hypothesis (i.e., H_0: α=0, β=1) is rejected if the F statistic exceeds the tabulated value of the F distribution with 2 and n-2 degrees of freedom. If this joint null hypothesis is rejected, separate t tests for bias and efficiency are desirable with the respective null hypotheses being α=0 (bias) and β=1 (inefficiency).

A visual inspection of the scatter plot often reveals the type of forecast bias. Figure 2 demonstrates the five basic types of linear results. The first type (upper right) is an under-forecast; the entire regression line is above the 45° line of perfect forecast. The second type is an over-forecast; the entire regression line is below the 45° line of perfect forecast. The third type is an over-forecast on low occupancy days and under-forecast on high occupancy days. The forth type is an under-forecast on low occupancy days and an over-forecast on high occupancy days. The fifth generic type is a perfect forecast.

The Mincer-Zarnowitz method provides a clear indication as to whether the forecast accuracy is

Figure 2
Regression line could be one of five basic types of errors.



acceptable and its results are statistically testable. The test can be programmed and is suitable for the task of forecast evaluations by the chain's corporate office.

Benchmarking assists

Forecasters often assess the quality of their prediction by comparing its accuracy to that of a competing set of predictions. This set of predictions serves as a benchmark model. If the forecast is not more accurate than the benchmark, the forecasting model should either be improved or replaced. The various approaches to benchmarking differ in two main respects: The type of benchmark used and the method of comparison.

Researchers advocate different types of benchmarks. Some suggest that a very simple forecast model would suffice, while other argue for the use of the most accurate forecast model available. A Naïve (simple) benchmark is very popular among scholars and practitioners.[15] The Naïve benchmark follows the following process: The next period's forecast is the most recently observed actual value. e.g., the forecast for tomorrow's occupancy is today's number of occupied rooms. f_{t+1} = At, where f_{t+1} is the forecast for period t+1, and A_t is the actual value observed at period t. A model with a trend[16] is given by: $A_t = A_{t-1} + M$. With seasonality the Naïve model is:

$$f_{t+i} = \frac{A_t(K_{j+s})}{K_j}$$

where t is the present time period, s is the number of periods ahead being forecasted and k is the adjustment index.

Among the advocates for a more sophisticated benchmark are Granger and Newbold[17] who suggest the Box-Jenkins model as the benchmark, Mincer and Zarnowitz who argue for the use of the best available extrapolative univariate method, and Newbold and Bos[18] who say, "It's important to evaluate forecast through comparison with the most serious alternatives that the analyst can find."

In practice, the nature of the evaluation task dictates the type of benchmark. A hotel that conducts an in-depth analysis of its occupancy forecast should compare the accuracy to that of the best available occupancy forecast method it can apply. The purpose of the in-depth analysis is to determine the appropriateness of a considered method. If the analyzed model is more accurate than the most accurate benchmark, it is a clear indication that it should be used. However, if the forecast model outperforms a simple Naïve model, the conclusion is not necessarily that the model is worth using. There might be another model that is more effective.

While this very same logic applies to the monitoring task of the corporate office, there are additional constraints to consider. More sophisticated benchmark models require a higher level of modeling skills and computing sophistication on the part of the monitoring unit. Consider, for example, the monthly forecast. To produce the "prediction" of a Naïve benchmark, one only needs up to 32 figures (the occupancy of the relevant month and the occupancy in the last day of the previous month to predict occupancy on the first day of the month), and, of course, the computation is very simple. The more sophisticated models are based on reservations figures; that means processing thousands of monthly figures for each hotel. Obviously, this large amount of information requires intensive computational power. Some models such as the Box-Jenkins even require judgment input in the process of forecasting. These requirements cannot be effectively met by the monitoring unit. The optimal benchmark for the monitoring task should be capable of producing the minimal acceptable accuracy. It should be fully automated and use a relatively small amount of information. Thus, for the monitoring task, the simple Naïve models seem more appropriate than the "stronger" benchmarks.

- **A simple ranking by the error:** A common practice among hospitality scholars[19] involves a simple ranking of the error measure(s). According to this approach, the error measure(s) of a forecast is compared to that of the competing set(s) of predictions. If the forecast is found to be more accurate than the benchmark (e.g., the forecast's 9.7 MSE is smaller than the benchmark's 10.3 MSE), then the forecast model is "declared" more accurate and therefore might be worth using. This approach is simple to apply, but it lacks a solid statistical justification, as no statistical test is applied to test the hypothesis that the forecast error is indeed significantly smaller than the benchmark.

- **Conditional efficiency:** Granger and Newbold[20] suggest that a set of predictions is conditionally efficient in respect to a second set of predictions (the benchmark) if a weighted average of both forecasts is not more accurate than the first one. In other words, the benchmark does not contain any useful information beyond that in the first set. The conditional efficiency can be statistically tested. Combining the two sets of predictions, f^1_t and f^2_t with w ($0 \leq w \leq 1$) one gets $(A_t - f^1_t) = w(f^2_t - f^1_t) + \varepsilon_t$. The hypothesis is H_0: w=0, H_1: w>0 and the OLS estimate of w is t tested.
- **Nonparametric methods:** The tests discussed so far require a normally distributed population, an assumption which is often invalid in reference to hotel occupancy rates. Moreover, it is unlikely that the distribution of the occupancy forecast error and its mean will remain constant over time. Such conditions, where the normality assumption is invalid, call for the use of "distribution free" non-parametric methods of analyzing forecast accuracy.

Researchers favor two nonparametric methods: the Sign test and the "stronger" Matched-Pairs Signed-Ranks test. For studies using the nonparametric tests in the general forecasting literature, see, for example, Flores[21] and Armstrong & Collopy.[22] For an application to hotel occupancy forecasting, see Schwartz and Hiemstra.[23] Daniel and Conover[24] describe these methods in details. The Sign test examines the percentage of times that the forecast is more accurate than the benchmark. This number is tested to be larger than .5 or 50 percent, using the binomial probability distribution or, if the sample is large enough, the normal approximation to the binomial distribution.

The Matched-Pairs Signed-Ranks test examines the sample of n values of differences, that is, for each forecasted day, the difference between the model's forecast error and the benchmark's forecast error (E^1_t- E^2_t) is calculated. The differences are ranked and the test statistic is based on the sum of ranks with positive signs. For a large sample,

$$Z = \frac{T - [n(n+1)]/4}{\sqrt{n(n+1)(2n+1)/24}}$$

is calculated for each pair of models compared, where T is the number of positive ranks. Z is distributed approximately as the standard normal. The hypotheses are: H_0: Md >= 0 and H_1: Md <0 where Md is the median of the population of differences. Wilcoxon's test assumes that the differences are independent, that they are measured on an interval scale or higher, and that the population is distributed symmetrically around the median. The advantage of Wilcoxon's test over the Sign test is that it takes into account the size of the error.

Both non-parametric methods are easily programmable and provide a solid statistical test. As such, they are excellent candidates for the evaluation of large number of forecasts, given that there is a set of alternative predictions with an acceptable level of accuracy that can be used as a benchmark.

Accuracy methods measured

For monitoring the accuracy of multiple occupancy forecasts, several methods are appropriate.

- **Error measure:** Since different hotels might have different cost of error functions, one cannot identify a single error measure that is likely to be more appropriate. All four types of error measures can be conveniently programmed. Thus, the main criterion in choosing a measure would be its ability to accurately reflect the damage caused by forecasting errors.
- **Evaluation method:** The Mincer-Zarnowitz efficiency test and Wilcoxon's non-parametric test emerge as the more adequate methods for monitoring the occupancy forecasts. Given a confidence level, the Mincer-Zarnowitz test assesses the forecast efficiency and the Wilcoxon test finds if the forecast is more accurate than an acceptable benchmark. Both are tests of statistical significance, both can be automated, and their results are easy to interpret. If feasible, it's recommended that both tests be performed. In the following example we demonstrate the use of the Mincer-Zarnowitz test and the Wilcoxon test.

The data (See Table 1) are taken from a 166-room hotel in the mid-west. It contains two sets of figures: the predicted and the actual daily occupancies for September 1996. The predicted figures are the combined product of a quantitative model and experts' predictions. Initial predictions were produced by the central reservation system at the corporate office using a back-propagation Neural-Networks algorithm. These predictions were reviewed and adjusted by the hotel's managers based on their experience and expectations (for more on Neural Networks see, for example, White.[26] The combination of model's predictions and managers' judgment in hotel occupancy forecasts is discussed in Schwartz.[27]) The adjusted forecast was then returned to the corporate office as the official monthly forecast report. The actual figures are taken from the variance report.

The use of the Mincer-Zarnowitz efficiency test and Wilcoxon's non-parametric test

Table 1
Actual daily occupancies, forecast, and various error measures

	Actual Occupancy	Forecast	Error	Absolute Deviation	Absolute Percentage Error	Squared Error	Asymmetric Error Measure
Sep-1-1996	120	120	0	0	0.0%	0	0
Sep-2-1996	95	90	-5	5	5.3%	25	100
Sep-3-1996	103	104	1	1	1.0%	1	40
Sep-4-1996	103	105	2	2	1.9%	4	80
Sep-5-1996	107	110	3	3	2.8%	9	120
Sep-6-1996	93	90	-3	3	3.2%	9	36
Sep-7-1996	105	92	-13	13	12.4%	169	676
Sep-8-1996	113	99	-14	14	12.4%	196	784
Sep-9-1996	59	65	6	6	10.2%	36	240
Sep-10-1996	53	60	7	7	13.2%	49	280
Sep-11-1996	27	55	28	28	103.7%	784	1120
Sep-12-1996	71	71	0	0	0.0%	0	0
Sep-13-1996	40	60	20	20	50.0%	400	800
Sep-14-1996	80	80	0	0	0.0%	0	0
Sep-15-1996	89	85	-4	4	4.5%	16	64
Sep-16-1996	118	101	-17	17	14.4%	289	1156
Sep-17-1996	122	97	-25	25	20.5%	625	2500
Sep-18-1996	104	97	-7	7	6.7%	49	196
Sep-19-1996	115	100	-15	15	13.0%	225	900
Sep-20-1996	89	85	-4	4	4.5%	16	64
Sep-21-1996	98	97	-1	1	1.0%	1	4
Sep-22-1996	115	102	-13	13	11.3%	169	676
Sep-23-1996	166	110	-56	56	33.7%	3136	12544
Sep-24-1996	88	86	-2	2	2.3%	4	16
Sep-25-1996	97	88	-9	9	9.3%	81	324
Sep-26-1996	113	90	-23	23	20.4%	529	2116
Sep-27-1996	143	110	-33	33	23.1%	1089	4356
Sep-28-1996	116	99	-17	17	14.7%	289	1156
Sep-29-1996	118	94	-24	24	20.3%	576	2304
Sep-30-1996	119	88	-31	31	26.1%	961	3844
			ME	MAD	MAPE	MSE	MASE
			-8	13	14.7%	325	1217

Figure 3
Line of perfect forecasts and regression line of actual vs. forecast

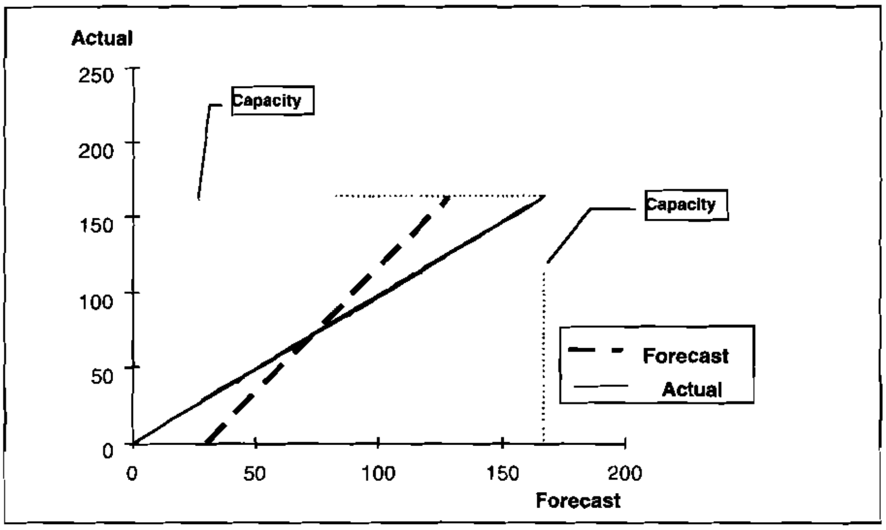


Table 2
Actual daily occupancies, Naïve forecast, and various error measures

	Actual Occupancy	Naïve	Error	Absolute Deviation	Absolute Percentage Error	Squared Error	Asymmetric Error Measure
Sep-1-1996	120	167	47	47	39.2%	2209	1880
Sep-2-1996	95	120	25	25	26.3%	625	1000
Sep-3-1996	103	95	-8	8	7.8%	64	256
Sep-4-1996	103	103	0	0	0.0%	0	0
Sep-5-1996	107	103	-4	4	3.7%	16	64
Sep-6-1996	93	107	14	14	15.1%	196	560
Sep-7-1996	105	93	-12	12	11.4%	144	576
Sep-8-1996	113	105	-8	8	7.1%	64	256
Sep-9-1996	59	113	54	54	91.5%	2916	2160
Sep-10-1996	53	59	6	6	11.3%	36	240
Sep-11-1996	27	53	26	26	96.3%	676	1040
Sep-12-1996	71	27	-44	44	62.0%	1936	7744
Sep-13-1996	40	71	31	31	77.5%	961	1240
Sep-14-1996	80	40	-40	40	50.0%	1600	6400
Sep-15-1996	89	80	-9	9	10.1%	81	324
Sep-16-1996	118	89	-29	29	24.6%	841	3364
Sep-17-1996	122	118	-4	4	3.3%	16	64
Sep-18-1996	104	122	18	18	17.3%	324	720
Sep-19-1996	115	104	-11	11	9.6%	121	484
Sep-20-1996	89	115	26	26	29.2%	676	1040
Sep-21-1996	98	89	-9	9	9.2%	81	324
Sep-22-1996	115	98	-17	17	14.8%	289	1156
Sep-23-1996	166	115	-51	51	30.7%	2601	10404
Sep-24-1996	88	166	78	78	88.6%	6084	3120
Sep-25-1996	97	88	-9	9	9.3%	81	324
Sep-26-1996	113	97	-16	16	14.2%	256	1024
Sep-27-1996	143	113	-30	30	21.0%	900	3600
Sep-28-1996	116	143	27	27	23.3%	729	1080
Sep-29-1996	118	116	-2	2	1.7%	4	16
Sep-30-1996	119	118	-1	1	0.8%	1	4

with the Naïve benchmark can be demonstrated. The following four error measures are used with the Wilcoxon test: Absolute Error, Absolute Percentage Error, Squared Error, and an Asymmetric Error Measure where the cost of under-forecasting error is squared and the cost of over-forecasting error is linear. The parameters of the Asymmetric function are set to ϕ=1 and φ=20.

- **The Mincer-Zarnowitz Efficiency Test:** Estimating $A_t = \alpha + \beta f_t + \varepsilon_t$, we get $\hat{\alpha} = 45.58$ and $\hat{\beta} = 1.59$. The hypothesis, H_0: α=0, β=1, is rejected since the statistic,

$$F = \left[\sum_{t=1}^{n}(A_t - f_t)^2 - \sum_{t=1}^{n}(A_t - a - bf_t)^2 \frac{1}{2}\right] \Big/ \left[\sum_{t=1}^{n}(A_t - a - bf_t)^2 \frac{1}{n-2}\right],$$

equals 12.80 and is larger than the tabulated value of F (for 95% confidence level), 3.34. Thus the test indicates that the forecast is inefficient.

Figure 3 shows that this set of predictions is of the third type where the higher the occupancy, the larger the under-forecast error.

- **Wilcoxon's Non-Parametric Test:** The Naïve Benchmark predictions along with four different error measures are given in Table 2.

Table 3
Wilcoxon test comparing forecast to Naïve, Random-Walk, Benchmark

Error measure	**Is forecast more accurate than the benchmark?** P(Z)	at 95% confidence level
Absolute deviation	.019	Yes
Squared error	.019	Yes
Absolute percentage error	.061	No
Asymmetric error	.827	No

The results of the Wilcoxon test are summarized in Table 3. Note the impact of the error measure on the test results. With the Absolute or Squared Deviation as the error measure, the forecast is found to be significantly more accurate than the Naïve benchmark, with a confidence level of 95 percent. When the Absolute Percentage or the Asymmetric Deviation are used as the error measure, the forecast is not significantly more accurate than the Naïve benchmark (with the same 95 percent confidence level).

Implications exist for managers

The Mincer-Zarnowitz efficiency test indicates that the forecast is inefficient. In this special case of a combined prediction (Neural Network forecast that has been adjusted by managers), the inaccuracy could be caused by the Neural Network (NN) forecasting model or by human bias. To identify the cause, one compares the accuracy of the original NN prediction to that of the adjusted prediction. Obviously, if the NN forecast is less accurate than the adjusted one, the NN model should be improved. Often, it means letting the same NN algorithm, backpropagation in our case, continue to "learn" the patterns until better accuracy is achieved with the hold-out sample. The problem could be due to "over training" in which case the Net needs to be re-trained. On rare occasions there is a need to apply a NN algorithm of a different type, for example, the hotel might replace the backpropagation model with a genetic NN algorithm.

When the NN forecast is more accurate than the adjusted one, it is likely that human bias is the reason for the inaccurate forecasts. Human bias, if proven to exist, can be better understood by studying Figure 3. Adjusted forecasts which are higher than 80 rooms tend to be lower than the actual, and adjusted forecasts that are lower than 70 rooms tend to be higher than the actual. Thus, if human bias exists, it is likely because managers tend to be over cautious when adjusting the predictions of the NN, over correcting high and low predictions. Hence, more accurate forecasts can be achieved if the hotel managers learn to accept the NN forecasts even when the numbers seem too high or too low.

With the Naïve model as a benchmark, Wilcoxon's non-parametric test leads to a different conclusion. When using AD, SE, and even APE as the error measures, the adjusted forecast is considered accurate and no correcting action by the hotel is required. This result emphasizes the importance of applying more than a single accuracy test. If both tests are applied, there is a better chance of identifying deteriorating accuracy.

Hotel chains often compare and rank the forecast accuracy of different hotels. It is important to understand that while this is a rather simple comparison, it is often misinterpreted. If one hotel produces more accurate forecasts it does not necessarily mean that a second hotel should adopt the first hotel's model. As accuracy depends on specific circumstances, adopting a model that performs better in a different location will not always improve the accuracy even after the model has been best fitted to the "adopting" hotel's data.

Monitoring the accuracy of hotels' occupancy forecasts is especially challenging as it requires that the method be both statistically sound and practical. Most important, the evaluation method must be programmable so it can monitor a large number of occupancy forecasts, and it must provide a clear and interpretable assessment of the quality of the assessed predictions.

Of the several basic types of error measures there is no universal error measure that can, *a priori*, be declared as the most appropriate for the task of monitoring the accuracy of occupancy forecasts. The error measure must reflect the cost of forecast error at the specific investigated hotel(s). If the evaluation process includes a comparison of accuracy across hotels of differing sizes, some type of standardization (e.g., a percentage-based error measure) is desirable.

Two evaluation methods seem particularly adequate for the task of occupancy forecast evaluation. From the group of "Absolute Worth" evaluation methods, the use of the Mincer-Zarnowitz efficiency test is recommended. Among the "Comparative" methods, the Wilcoxon Matched-Pairs Signed-Ranks test and the Naïve benchmark are most appropriate for the monitoring task.

Appendix A

Common Error Measures

Mean Absolute Deviation, MAD	$\frac{1}{n}\sum_{t=1}^{n}\lvert E_t\rvert$
Sum of Squared Error, SSE	$\sum_{t=1}^{n} E_t^2$
Mean Squared Error, MSE	$\frac{1}{n}\sum_{t=1}^{n} E_t^2$
Mean Absolute Percentage Error, MAPE	$\frac{1}{n}\sum_{t=1}^{n}\frac{\lvert E_t\rvert}{A_t}$
Standard Deviation of Error, SDE	$\sqrt{\sum_{t=1}^{n} E_t^2 \frac{1}{n-1}}$

References

[1] E. Nebel, *Managing Hotels Effectively - Lessons From Outstanding General Managers* (New York: Van Nostrand Reinhold, 1991).

[2] A. De Franco, "The Importance and Use of Financial Forecasting and Budgeting at The Departmental Level in the Hotel Industry as Perceived by Hotel Controllers," *Hospitality Research Journal* 20, no. 3 (1997): 99-110; R. Schmidgall, "Forecasting For Profitability," *Bottomline* 4, No. 2 (1989): 20-23.

[3] M. Coltman, *Hospitality Management Accounting* (New York: Van Nostrand Reinhold, 1994); Nebel.

[4] R. Fildes, "The Evaluation of Extrapolative Forecasting Methods," *International Journal of Forecasting* 8 (1992): 81-98; S. Wheelwright and S. Makridakis, *Forecasting Methods For Management* (New York: John Wiley & Sons, 1989).

[5] J. Miller, C. McCahon, and J. Miller, "Foodservice Forecasting Using Simple Mathematical Models," *Hospitality Research Journal* 15, No. 1 (1990): 43-58; J. Miller, C. McCahon, and J. Miller, "Foodservice Forecasting: Differences in Selection of Simple Mathematical Models Based on Short-Term and Long-Term Data Sets," *Hospitality Research Journal* 16, No. 2 (1993): 95-102.

[6] W. Andrew, D. Cranage and C. Lee, "Forecasting Hotel Occupancy Rates with Time Series Models: an Empirical Analysis," *Hospitality Research Journal* 14, No. 2 (1991): 173-180.

[7] D. Cranage and W. Andrew, "A Comparison of Time Series and Econometric Models for Forecasting Restaurant Sales," *International Journal of Hospitality Management* 11, No. 2 (1002): 129-142.

[8] A. Messersmith and J. Miller, *"Forecasting in Foodservice"* (New York: John Wiley & Sons, 1992).

[9] Z. Schwartz, "Convention and Conference Facilities: A Framework of Statistical Predictions and Judgemental Adjustments for Daily Occupancy Forecasts," *Journal of Convention and Exhibition Management* 1, No. 1 (1997): 71-88.

[10] Z. Schwartz and S. Hiemstra, "Improving the Accuracy of Hotel Reservations Forecasting: Curves Similarity Approach," *Journal of Travel Research* 36, No. 1 (1997): 3-14.

[11] H. Theil, *Economic Forecasts and Policy* (Amsterdam: North Holland, 1958).

[12] G. Janacek, "Estimation of Minimum Mean Squared Error of Prediction," *Biometrika* 62 (1975): 175-180.

[13] A. Kolmogorov, "Stationary Sequences in Hilbert Space (in Russian)," *Bull. Math. Univ. Moscow* 2, No. 6 (1941): 1-40.

[14] J. Mincer and V. Zarnowitz, "The Evaluation of Economic Forecasts," in Mincer, J., (editor), *Economic Forecasts and Expectations* (New York: National Bureau of Economic Research, 1969).

[15] Miller, et al.

[16] Wheelwright and Makridakis.

[17] C. Granger and P. Newbold, "Some Comments on the Evaluation of Economic Forecasts," *Applied Economics* 5 (1973): 35-47.

[18] P. Newbold and T. Bos, *Introductory Business and Economic Forecasting* (Cincinnati, Ohio: South-Western Publishing Co., 1994).

[19] Andrew, et al.; Cranage and Andrew; Miller, et al.

[20] C. Granger and P. Newbold, *Forecasting Economic Time Series* (San Diego, Calif.: Academic Press, Inc., 1986).

[21] B. Flores, "Use of the Sign Test to Supplement the Percentage Better Statistic," *International Journal of Forecasting* 2, No. 4 (1986): 477-489.

[22] J. Armstrong and F. Collopy, "Rule-Based Forecasting: Development and Validation of an Expert Systems Approach to Combine Time Series Extrapolations," *Management Science* 38, No. 10 (1990): 1394-1412.

[23] Schwartz and Hiemstra.

[24] W. Daniel, *Applied Nonparametric Statistics* (Boston: PWS-KENT, 1990); W. Conover, *Practical Nonparametric Statistics* (New York: John Wiley & Sons, 1980).

[25] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics* 1 (1945): 80-83.

[26] H. White, et al., *Artificial Neural Networks: Approximation and Learning Theory* (Cambridge, Mass.: Blackwell Publishers, 1992).

[27] Schwartz.

***Zvi Schwartz** is a lecturer in the Department of Hospitality and Tourism in the School of Management at Ben-Gurion University of the Negev, Israel.*