AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

*The World's Largest Open Access Agricultural & Applied Economics Digital Library*

**This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.**

**Help ensure our sustainability.**

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

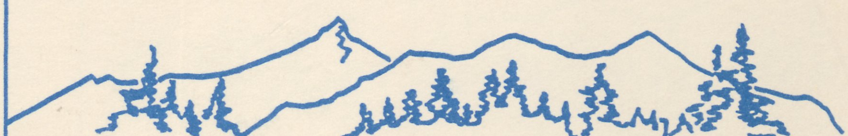
*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

*No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.*

338.1
G67
1991

# Papers of the 1991 Annual Meeting

# Western Agricultural Economics Association



Waite Library
Dept. Of Applied Economics
University of Minnesota
1994 Buford Ave - 232 ClaOff
St. Paul MN 55108-6040

Portland, Oregon
July 7 – 10, 1991

**Qualitative Forecast Evaluation: A Comparison of Two Performance Measures**

Christopher S. McIntosh, Univ. of GA
Jeffrey H. Dorfman, Univ. of GA

*The standard turning point analysis using ratios of successful to unsuccessful predictions is compared to nonparametric test of the null hypothesis of no information on the direction of price movements. The inability of the turning point ratio method to correctly evaluate series with a trend is demonstrated.*

Agricultural producers and marketers often have access to several sets of forecasts on the same series. In addition, several methods for combining sets of forecasts are available. Various well established methods exist for measuring the quantitative accuracy of point estimate forecasts. In contrast, there is less agreement on how to measure the qualitative accuracy of a set of forecasts; that is, the forecast's ability to predict whether the series will move in an upward or downward direction from its present value. The ability to predict the direction a series will move is non-trivial since for asset price series and markets for futures and options an accurate forecast of the direction a series will move is at least as important as a forecast of the expected new level. In these markets, money can be made simply by knowing (guessing) the direction in which the series will move.

Researchers have recognized the importance of forecasting the direction a series will move and have increasingly devoted resources to the study of the construction and evaluation of qualitative forecasts (e.g. Feather and Kaylen). Some research has concentrated on the prediction of turning points themselves (see, for example, Kling; Zellner et al.), others on evaluating a forecasts ability to predict the direction of revision (change) in a series (see, for example, Bessler and Brandt; Brandt and Bessler 1981, 1983; Harris and Leuthold; Kost; Naik and Leuthold)[1].

This paper compares a success rate measure of qualitative forecast ability due to Naik and Leuthold to a probability based measure developed by Henriksson and Merton. We show that the Henriksson-Merton measure provides additional, and more accurate, information concerning a series' qualitative forecasting ability than provided by a simple success rate measure. Although the Henriksson-Merton test is well known in the business literature as a test of market timing ability, it has received almost no attention in the agricultural economics or forecasting literature. A single exception to this is found in an analysis of U.S. cattle prices using multivariate time series models (Park).

## Definitions

Consider the case of one-step-ahead forecasts, a definition that has appeared in recent literature (Zellner et al.; Kaylen and Brandt) for the qualitative features of a one-step-ahead forecast of a turning point at period $t$-1 is as follows:

(1a) $X_{t-2} < X_{t-1}$ and $\hat{X}_t < X_{t-1}$ $\equiv$ peak turning point (PTP),

(1b) $X_{t-2} < X_{t-1}$ and $\hat{X}_t > X_{t-1}$ $\equiv$ upward non-turning point (UNTP),

(1c) $X_{t-2} > X_{t-1}$ and $\hat{X}_t < X_{t-1}$ $\equiv$ downward non-turning point (DNTP), (1d) $X_{t-2} > X_{t-1}$ and $\hat{X}_t > X_{t-1}$ ≡ trough turning point (TTP),

where the $X_t$ are realizations of some series at time $t$ and the $\hat{X}_t$ are the forecast of $X_t$ made at time $t$-1. The ability of a forecast to correctly predict the events described above is often used as a measure of qualitative forecast performance.

The forecast of the direction of revision in a series can be defined as follows:

(2a) $\hat{X}_t < X_{t-1}$ ≡ downward movement

(2b) $\hat{X}_t > X_{t-1}$ ≡ upward movement.

This collapses the 4 types of revision defined in (1a) through (1d) into two types of revision.

## Success Rate Measures

Bessler and Brandt classified forecasted versus actual turning points by examining whether or not a change in direction of the series occurred when it had been forecast. This method relies on the definitions presented in equations 2a and 2b. They construct a 2 x 2 contingency table of the change and no change realizations.

Naik and Leuthold later developed a 4 x 4 contingency table and a set of related summary measures based on a definition of a turning point similar to that in equations (1a) through (1d). The summary measures developed by Naik and Leuthold include the ratio of accurate forecasts (RAF) the ratio of inaccurate forecasts (RIF), the ratio of worst forecasts (RWF), and the ratio of accurate to worst forecasts (RAWF)[1]. The RAF statistic is calculated as the sum of correct turning point forecasts (including UNTP and DNTP) divided by the sum of all forecasts. The RIF statistic is calculated as the sum of the number times a PTP or TTP is forecast and an UNTP or DNTP occurs and the number of times a UNTP or DNTP is forecast and a PTP or TTP occurs, divided by the total number of forecasts (or more simply, 1-RAF). Although the RAF is a numerical measure, it only provides an ordinal ranking of competing forecasts. There is no way of knowing, from the RAF measure alone, if a value of 0.68 is "good" or how much better 0.78 is than 0.75.

If the decision maker is interested in the direction of revision of a price series, this information is contained (conditioned on past information) in the RAF measure. Note that although Naik and Leuthold used a 4 x 4 contingency table instead of Bessler and Brandt's 2 x 2 table, the RAF measure makes no distinction between turning points and non-turning points. The RAF measure simply indicates the percentage of correct direction of revision forecasts out of the total number of forecasts. What the RAF doesn't reflect, however, is how the number of correct directional forecasts is divided between correct upturns and correct downturns. Again, the ability to correctly predict whether price will move up or down is important when dealing with series where monetary gains and losses are directly linked to upward and downward movements in the series (Cerchi and Havenner). Consequently the RAF measure fails to tell "the rest of the story" and may provide misleading information about the quality or information content of a forecast. For example, if an actual series consistently went up-up-down-up-up-down-up-up-down- ... and a set of forecasts simply predicted that prices would go up, it would achieve an RAF of 0.667, but it obviously contains little information regarding price movements.

## A Probability Measure

If a set of forecasts is capable of predicting the direction of revision of a series, then it provides valuable information to decision makers. Henriksson and Merton developed a nonparametric statistic to measure the ability of a set of forecasts in predicting the direction of revision. This test is different form all other measures of qualitative accuracy commonly found in the agricultural economics literature in that its scale is a probability level. This allows an exact and meaningful interpretation to be placed on the performance differences between two sets of forecasts. While it is unclear whether two RAFs of .750 and .785 are significantly different, a difference in probability level from (for example) .560 to .595 is easily understood and interpreted.

The Henriksson-Merton test (hereafter, HM test) is based on a formal hypothesis test for information value in a set of forecast predictions on direction of revision. The null hypothesis of the finite sample test is that the forecast has no information value (i.e. cannot predict the direction of revision in the series). Because the test produces a statistic with a known distribution (without any distributional assumptions concerning the forecasts themselves), the HM test can provide exact confidence and significance levels for the rejection of the hypothesis of no information.

Let $\gamma_t = 1$ if the forecast made at time t-1 for time t is that $\hat{X}_t > X_{t-1}$ and $\gamma_t = 0$ if the forecast is that $Xt \leq X_{t-1}$. The probabilities for $\gamma_t$ conditional on the realized outcome can be written as:

$$\text{(3a)} \quad \text{prob} \{ \gamma_t = 0 \mid X_t \leq X_{t-1} \} \equiv p_1,$$

and

$$\text{(3b)} \quad \text{prob} \{ \gamma_t = 1 \mid X_t > X_{t-1} \} \equiv p_2.$$

Therefore, $p_1$ is the conditional probability of a correct forecast of a downward revision and $p_2$ is the conditional probability of a correct forecast of an upward revision. It is assumed that $p_1$ and $p_2$ do not depend on the magnitude of $(X_t - X_{t-1})$. Hence the conditional probability of a correct forecast depends only on whether $X_t$ is greater than or less than $X_{t-1}$.

Merton shows that a necessary and sufficient condition for a set of forecasts to have no information value is $p_1 + p_2 = 1$, and that a necessary condition for the forecasts to have a positive information value is $p_1 + p_2 > 1$. For example a perfect forecaster will have $p_1 = 1$ and $p_2 = 1$, and therefore $p_1 + p_2 = 2 > 1$. Similarly forecasts with $p_1 + p_2 < 1$ can be shown to have a negative information value because they are systematically incorrect. However, a systematically incorrect forecast is perverse in the sense that $p_1' \equiv 1 - p_1$ and $p_2' \equiv 1 - p_2$ would satisfy $p_1' + p_2' > 1$ and, therefore, have a positive information value provided that one is aware that the forecasts are perverse.

Consequently, to test the information value of a forecast one could examine whether $p_1 + p_2 = 1$ (i.e. the null hypothesis of no information value is $H_o$: $p_1 + p_2 = 1$). Because $\gamma_t$ is a binary random variable, an exact finite sample distribution for the test statistic $(p_1 + p_2)$ can be derived using the hypergeometric distribution. This allows the calculation of an exact confidence level for the rejection of the null hypothesis. The test is only concerned with the $\gamma_t$'s and not the levels of the forecasts or the parameters of the model used to generate the forecasts. Therefore, the test has finite sample properties even when used to evaluate forecasts from time series models in which the parameters have only asymptotic properties. To calculate the confidence level, $c$, of rejecting the null hypothesis of no information, define for each series the following values over the forecasts being tested:

$N_1$ = the number of observations with downward movement
$N_2$ = the number of observations with non-downward movement
$N = N_1 + N_2$
$n_1$ = the number of correct forecasts of downward movement
$n_2$ = the number of incorrect forecasts of downward movement
$n = n_1 + n_2$ = the number of forecasts of downward movement

The confidence level, $c$, is given by the formula

$$(4) \qquad c = 1 - \sum_{x=n_1}^{\min(N_1,n)} \binom{N_1}{x}\binom{N_2}{n-x} \Big/ \binom{N}{n}.$$

If the forecast series are not perverse, then it would never be true that $p_1 + p_2 < 1$ and thus a one-tailed test is suggested. For a one-tail test with a confidence level of $c$, the null hypothesis of no information value would be rejected at any level of significance greater than $1 - c$. For additional details on the construction of this test see Merton and Henriksson and Merton.

## Comparing the Measures

Table 1 shows the number of times a direction of revision was forecast correctly for seven sets of forecasts on hog prices. The sets of forecasts examined were originally published in Brandt and Bessler (1981), and re-examined by Naik and Leuthold and Kaylen and Brandt. The forecasts are shown, for reference, in table 2.

Table 1 also shows the values for the HM test and the RAF for all seven sets of hog price forecasts. Using the definitions of turning points given in (1a) through (1d) above (and in Kaylen and Brandt), and the RAF measure of forecast quality suggested by Naik and Leuthold, one could conclude that the two-period adaptive composite, the simple average of two forecasts and the simple average of three forecasts are of equal qualitative value. Each of these forecasts achieved a ratio of accurate forecasts of 0.75. Evaluating the two-period adaptive composite and the simple average of two forecasts using the HM test, one would fail to reject the null hypothesis of no information value at the 0.10 level (the significance level for both forecast series is 0.152). The null hypothesis would be rejected (at the 0.071) level for the simple average of three forecasts. This statistic reflects the fact that the simple average of three forecasts correctly predicted the direction of revision in a more balanced manner than the other forecasts (i.e. it was successful at predicting price decreases as well as increases). Similarly the econometric, expert opinion and minimum variance forecasts might be considered equivalent based on the ratio of accurate forecasts, which was 0.667 for all three. However, we would reject the null hypothesis of no information value at the 0.406 level for the econometric model, while rejecting the null at the 0.273 level for the minimum variance and 0.141 for the expert opinion forecasts.

As a further example, using the set of actual hog prices in table 2, a set of forecasts that predicted upward movement for every revision would achieve and RAF of 0.750 (because 8 of the 12 revisions were upward), but the HM confidence level would be 0.00. Such a set of forecasts would have achieved an RAF equal to the highest RAFs in table 1, but its HM confidence level would be lower than any of the forecasts in table 1 (i.e. 0.00).

This again highlights the fact that for the RAF measure only the percentage of correct directional forecasts matters. The HM test takes into account separately the ability to predict upward and downward revisions, as well as accounting for a series where the number of upturns and downturns is not balanced. This is an important feature in a measure of qualitative forecast accuracy when the series being forecast have trends (such as many economic time series). In a series with a preponderance of either upturns of downturns a naive model (such as always forecasting up) could achieve a high RAF. In contrast, such a model would have a low (or zero) HM confidence level.

The HM test, because it accounts for nonequal numbers of upward and downward revisions, provides a correct basis of comparison for all series. The value assigned to forecasting results that are equivalent to what could be expected by pure chance will always be a confidence level of $c = 0.0$, regardless of the actual "success rate." The RAF does not provide such a basis for comparison. The RAF level attributable to pure chance varies by series according to the relative frequency of upward and downward revisions.

## Concluding Remarks

The 4 x 4 contingency table and associated measures developed by Naik and Leuthold add an important tool to the array of forecast evaluation metrics. However, a measure of qualitative accuracy is needed that can provide more than a simple ordinal ranking and thus reflect information that the 4 X 4 contingency table cannot. In addition, the RAF measure suggested by Naik and Luethold may give very misleading results if the forecast and actual series are trending up or down.

The test demonstrated here provides an additional metric to judge the qualitative accuracy of forecasts where ability to predict direction of revision is important. Further, because the test has a formally stated null hypothesis and a known finite sample distribution, it provides the decision maker with a statistically based measure of the qualitative performance of a set of forecasts. The level of confidence indicated by this test is accurate even if a trend is present in the forecast and actual series.

## Footnotes

[1] Naik and Leuthold provide a good discussion of the literature on measures for qualitative forecast evaluation

[2] For detailed definitions of these measures see Naik and Leuthold.

## References


Bessler, D. A. and J. A. Brandt. "Forecasting Livestock Prices with Individual and Composite Methods." *Applied Economics* 13(1981):513-22.

Brandt, J. A., and Bessler, D. A. "Composite forecasting: an application with U.S. hog prices." *American Journal of Agricultural Economics*, 63(1981):135-40.

Brandt, J. A., and Bessler, D. A. "Forecasting with Vector Autoregression versus Univariate ARIMA Process: An Empirical Example with U.S. Hog Prices." *North Central Journal of Agricultural Economics*, 2(1983):237-48.

Cerchi, M. and A. Havenner. "Cointegration and Stock Prices, The Random Walk on Wall Street Revisited." *Journal of Economic Dynamics and Control* 12(1988):333-46.

Feather, P. M. and M. S. Kaylen. "Conditional Qualitative Forecasting." *American Journal of Agricultural Economics* 71(1989):195-201.

Harris, K. S. and R. M. Leuthold. "Alternative Forecasting Techniques: A Case Study for Livestock Prices." *North Central Journal of Agricultural Economics* 7(1985):40-50.

Henriksson, R. D. and R. C. Merton. "On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Forecasting Skills." *Journal of Business* 54(1981):513-33.

Kaylen, M. S. and J. A. Brandt. "A Note on Qualitative Forecast Evaluation: Comment." *American Journal of Agricultural Economics* 70(1988):415-16.

Kling, J. L. "Predicting the Turning Points of Business and Economic Time Series." *Journal of Business* 60(1987):201-38.

Kost, W. "Model Validation and the Net Trade Model." *Agricultural Economics Research*, 32(1980):1-16.

Merton, R. C. "On Market Timing and Investment Performance. I. An Equilibrium Theory of Value for Market Forecasts." *Journal of Business* 54(1981):363-406.

Naik, G. and R. M. Leuthold. "A Note on Qualitative Forecast Evaluation." *American Journal of Agricultural Economics* 68(1986):721-26.

Park, T. "Forecast Evaluation for Multivariate Time-Series Models: The U.S. Cattle Market." *Western Journal of Agricultural Economics* 15(1990):133-43.

Zellner, A., C. Hong, and G. M. Gulati. "Turning Points in Economic Time Series, Loss Structures and Bayesian Forecasting." in *Bayesian and Likelihood Methods in Statistics and Econometrics*, Geisser, Hodges, Press and Zellner eds. North Holland, 1990.

Table 1. Turning Points and a Test for Value of Information.

| Forecast[b] | Turning Point Analysis[a] Ratio of Accurate Forecasts[c] | Ratio of Inaccurate Forecasts[c] | Number Correct Upturns[d] | Number Correct Downturns[d] | HM Confidence Level[e] |
|---|---|---|---|---|---|
| Econometric | 0.667 | 0.333 | 6 | 2 | 0.594 |
| ARIMA | 0.583 | 0.417 | 4 | 3 | 0.576 |
| Expert Opinion | 0.667 | 0.333 | 4 | 4 | 0.859 |
| Two-Period Adaptive | 0.750 | 0.250 | 6 | 3 | 0.848 |
| Minimum Variance | 0.677 | 0.333 | 5 | 3 | 0.727 |
| Simple Average (2)[f] | 0.750 | 0.250 | 6 | 3 | 0.848 |
| Simple Average (3)[g] | 0.750 | 0.250 | 5 | 4 | 0.929 |

[a] Turning points are defined in equations (1a-d).

[b] Forecasts are from Brandt and Bessler (1981).

[c] The ratio of accurate forecasts and ratio of inaccurate forecasts are defined in Naik and Leuthold. Because none of these forecasts were "worst" forecasts as defined by Naik and Leuthold the ratio of worst forecasts and ratio of accurate to worst forecasts were not calculated.

[d] The total number of upturns in the actual series was 8, the number of downturns was 4. Upturns and downturns are defined in equations (2a-b). Note that the initial upturn was discarded in calculating these totals, as well as the HM test results, so that the number of upturns and downturns here would equal those from the turning point analysis (RAFs).

[e] (1 - the Henriksson-Merton confidence level) gives the significance level, i.e. the highest level at which one would fail to reject the null hypothesis of no information ($H_o$: $p_1 + p_2 = 1$) using a one-tailed test.

[f] Composite of Econometric and ARIMA models

[g] Composite of Econometric, ARIMA and expert opinion models.