"EVALUATING ACCURACY (OR ERROR) MEASURES"

by

S. MAKRIDAKIS*
and
M. HIBON**

95/18/TM

* Research Professor of Decision Sciences and Information Systems at INSEAD, Boulevard de Constance, Fontainebleau 77305 Cedex, France.

** Research Associate at INSEAD, Boulevard de Constance, Fontainebleau 77305 Cedex, France.

EVALUATING ACCURACY (OR ERROR) MEASURES

Spyros Makridakis and Michèle Hibon

INSEAD

Abstract

This paper surveys all major accuracy measures found in the field of forecasting and evaluates them according to two statistical and two user oriented criteria. It is established that all accuracy measures are unique and that no single measure is superior to all others. Instead there are tradeoffs in the various criteria that must be considered when selecting an accuracy measure for reporting the results of forecasting methods and/or comparing the performance of such methods. It is concluded that symmetric MAPE and Mean Square Error are to be preferred for reporting or using the results of a specific forecasting method while the difference between the MAPE of NAIVE 2 minus that of a specific method is a prefereable way of evaluating some specific method to some appropriate benchmark.

EVALUATING ACCURACY (OR ERROR) MEASURES

Spyros Makridakis and Michèle Hibon

INSEAD

The purpose of this paper is to study accuracy (or error) measures from both a statistical and practical point of view. Such measures are indispensable for helping us to (a) choose an appropriate model among the many available, (b) select the best method for our particular forecasting situation, (c) measure (and report) the most likely size of forecasting errors and (d) quantify (and report) the extent of uncertainty surrounding the forecasts. In judging the value of models/methods and the size of their forecasting errors/uncertainty, it is critical to distinguish between model fitting and post-sample (referring to periods beyond those for which historical data is available and used for developing the forecasting model) measures. Research has shown that post-sample accuracies are not always related to those of the model that best fits available historical data (Makridakis, 1986; Pant and Starbuck, 1990). The correlations between the two are small to start with (0.22 for the first forecasting horizon) and become equal to zero for horizons longer than four periods ahead. This means that we must judge the appropriateness of whichever measure we use by how effectively it provides information about post-sample performance.

For post-sample comparisons, research findings indicate that the performance (accuracy) of different methods depends upon the accuracy measure used (reference). This means that some methods are better when, for example, Mean Absolute Percentage Errors (MAPEs) are used while others are better when rankings are utilized, although the various accuracy measures are clearly correlated (Armstrong and Collopy, 1992). From a theoretical point of view there is a problem as no single method can be designated as the 'best' (see Winkler and Murphy, 1992), although there might be methods that perform badly in all accuracy measures. From a practical point of view the 'best' accuracy measure has to be related to the purpose of forecasting, its value for improving decision making, and the specific needs and concerns of the person or situation using the forecasts. Thus, in a one-time auction the method that comes up the best most of the time is to be preferred (e.g., the percentage better measure) while in repeated auctions average ranks should be selected as in both cases the size of forecasting errors is of no importance. In budgeting the MAPE may be most appropriate as it conveys information about average percentage errors which are used in reporting accounting results and profits. In inventory situations, on the other hand, Mean Square Errors (MSE) are the most relevant as a large error is much less desirable than two or more smaller ones whose sum is about the same as the large error. Finally, in empirical comparisons when many objectives are to be satisfied at once, several (or even many) measures may have to be used.

Is there a best overall measure that can be used in the great majority of situations and which satisfies both theoretical and practical concerns? Surprisingly, very little objective evidence exists to answer such a question. To this end the work of Armstrong and Collopy (1992) as well as Fildes (1992) are important contributions in both raising, once more, the issue of what constitutes the most appropriate measure and in providing objective information to judge the advantages/drawbacks of such measures.

Our paper is organized in three sections. First, a brief description of all major accuracy (or error) measures found in the forecasting literature is provided together with a discussion of the advantages and drawbacks of each. Second, four criteria (two statistical and two user oriented) are presented and the various accuracy measures are evaluated in terms of each. A major conclusion of this paper is that it is not possible to optimize these criteria at the same time, as achieving one requires a tradeoff in another. Third, there is a discussion and directions for future research section where the various accuracy measures are compared and a new way of evaluating methods is suggested. It is concluded that the MSE is the most appropriate measure for selecting an appropriate forecasting model while the MAPE (symmetric) is the most appropriate measure for evaluating the errors of single series **and** making meaningful comparisons across many series. In addition the difference of the APE (Absolute Percentage Error) of a specific method from Naive 2 (Deseasonalized Random Walk) is, in our view, the most appropriate way of making benchmark comparisons than alternatives such as Theil's U-Statistic. It is also concluded that the MSE is the most useful, and only way, of measuring the uncertainty in the forecasts and using it for determining optimal levels of stocks in inventory models. For future research directions it is shown how various measures can be compared to Naive 2 (a deseasonalized random walk) by estimating their beta coefficients when a regression model is run with the independent variable, the error measure of Naive 2, and the dependent variable, the same error measure of the forecasting method we are interested in. We believe that this is a major avenue for further research.

1. ACCURACY MEASURES: BRIEF DESCRIPTION AND DISCUSSION

There are fourteen accuracy measures which can be identified in the forecasting literature. A brief description and discussion of each is provided next.

1.1. Mean Square Error (MSE)

The mean square error is defined as follows:

$$\mathrm{MSE} = \frac{\Sigma(X_t - F_t)^2}{m} = \frac{\Sigma e_t^2}{m} \tag{1}$$

where X_t is the actual data at period t

F_t is the forecast (using some model/method) at period t

e_t is the forecast error at period t

while m is the number of methods (or observations) used in computing the MSE.

For the purpose of comparing various methods the summation goes from 1 to m, where m is the total number of series summed up at period t to compute their average.

For the purpose of evaluating the post-sample accuracy of a single method m is the total number of periods available for making such an evaluation. Alternatively, m can denote the number of observations (historical data) available and used to determine the best model to be fitted to such data.

The MSE, as its name implies, provides for a quadratic loss function as it squares and subsequently averages the various errors. Such squaring gives considerably more weight to large errors than smaller ones (e.g., the square error of 100 is 10000 while that of 50 and 50 is only 2500 + 2500 = 5000, that is half). MSE is, therefore, useful when we are concerned about large errors whose negative consequences are proportionately much bigger than equivalent smaller ones (e.g., a large error of 100 vs two smaller ones of 50 each).

Expression (1) is similar to the statistical measure of variance which allows us to measure the uncertainty around our most likely forecast F_t. As such the MSE plays an additional, equally important role, in allowing us to know the uncertainty around the most likely predictions which is a prerequisite to determine optimal inventory levels.

An alternative way of expressing the MSE is by computing the square root of expression (1), or

$$\text{Root Mean Square Error (RMSE)} = \sqrt{\frac{\Sigma(X_t - F_t)^2}{m}} = \sqrt{\frac{\Sigma e_t^2}{m}} \tag{2}$$

The range of values that expression (1) or (2) can take is from 0 to $+\infty$ making comparisons among series, or time horizons difficult and raising the prospect of outliers that can unduly influence the average computed in (1) or (2). Expression (2) is similar to the standard deviation also used widely in statistics.

The two biggest advantages of MSE or RMSE are that they provide a quadratic loss function and that they are also measures of the uncertainty in forecasting. Their two biggest disadvantages are that they are absolute measures that make comparisons across forecasting horizons and methods highly problematic as they are influenced a great deal by extreme values. Chatfield (1988), for instance, concluded that a small number (five) out of the 1001 series of the M-Competition determined the value of RMSE because of their extreme errors while the remaining 996 had much less impact. The MSE is equally used by academicians (see Zellner, 1986) and practitioners (see Armstrong and Carbone, 1982).

1.2. Mean Absolute Error (MAE)

The mean absolute error is defined as

$$\text{MAE} = \frac{\Sigma|X_t - F_t|}{m} = \frac{\Sigma|e_t|}{m} \tag{3}$$

when $|X_t - F_t|$ means absolute (ignoring negative signs, i.e., negative errors) value.

The MAE is also an absolute measure like the MSE and this is its biggest disadvantage. Its value fluctuates from 0 to $+\infty$. However, since it is not of quadratic nature, like the MSE, it is influenced less by outliers. Furthermore, because it is a linear measure its meaning is more intuitive; it tells us about the average size of forecasting errors when negative signs are ignored. The biggest advantage of MAE is that it can be used as a substitute for MSE for determining optimal inventory levels (see Brown, 1962). The MAE is not used much by either practitioners or academicians.

1.3. Mean Absolute Percentage Error ($MAPE_{reg}$)

The mean absolute percentage error is defined as

$$\mathrm{MAPE} = \frac{\Sigma\left|\frac{X_t - F_t}{X_t}\right|}{m}(100) = \frac{\Sigma\left|\frac{e_t}{X_t}\right|}{m}(100) \qquad (4)$$

The MAPE is a relative measure which expresses errors as a percentage of the actual data. This is its biggest advantage as it provides an easy and intuitive way of judging the extent, or importance of errors. In this respect an error of 10 when the actual value is 100 (making a 10% error) is more worrying than an error of 10 when the actual value is 500 (making a 2% error). Moreover, percentage errors are part of the everyday language (we read or hear that the GNP was underestimated by 1% or that unemployment increased by 0.2% etc) making them easily and intuitively interpretable. Furthermore, because they are relative they allow us to average them 'across' forecasting horizons and series. In addition we can make comparisons involving more than one method since the MAPE of each tells us about the average relative size of their errors. Such averaging across horizons and or methods makes much more sense than doing so with MSE or, in this respect, with practically all other error measures described below.

MAPE is used a great deal by both academicians and practitioners and it is the only measure appropriate for evaluating budget forecasts and similar variables whose outcome depends upon the proportional size of errors relative to the actual data (e.g., we read or hear that the sales of company X increased by 3% over the same quarter a year ago, or that actual earnings per share were 10% below expectations).

The two biggest disadvantages of MAPE are that it lacks a statistical theory (similar to that available for the MSE) on which to base itself and that equal errors when X_t is larger than F_t give smaller percentage errors than when X_t is smaller than F_t. For instance, when the actual value, X_t, is 150 and the forecast, F_t, is 100, the Absolute Percentage Error (APE) is:

$$\text{APE}_t\left|\frac{X_t - F_t}{X_t}\right| = \frac{150-100}{150} = \frac{50}{150} = 33.33\%$$

However, when $X_t = 100$ and $F_t = 150$ (still resulting in an absolute error of 50) the APE is:

$$\text{APE} = \left|\frac{100-150}{100}\right| = \frac{50}{100} = 50\%$$

This difference in absolute percentage errors when $X_t > F_t$ vs $X_t < F_t$ can create serious problems when the value of X_t is small (close to zero) and F_t is big, as the size of the APE can become extremely large making the comparisons among horizons and/or series sometimes meaningless. Thus the MAPE can be influenced a great deal by outliers as its value can become extremely large (see below).

1.4. The Symmetric Mean Absolute Percentage Error (MAPE_{sym})

The problem of asymmetry of MAPE and its possible influence by outliers can be corrected by dividing the forecasting error, e_t, by the average of both X_t and F_t, or

$$\text{APE}_t = \left|\frac{X_t - F_t}{(X_t + F_t)/2}\right|(100) \tag{5}$$

Using expression (5) will yield an APE of 40% whether $X_t = 150$ and $F_t = 100$, or $X_t = 100$ and $F_t = 150$, as it requires dividing the error 50 by the average of $X_t + F_t$ which is 125 in both cases.

We will call the MAPE found by expression (5) symmetric as it does not depend on whether X_t is higher than F_t or vice versa, while we will refer to the MAPE of expression (4) as regular. Or

$$\text{MAPE}_{\text{sym}} = \Sigma\left|\frac{X_t - F_t}{(X_t - F_t)/2}\right| / \, m(100) \tag{6}$$

while

$$\text{MAPE}_{\text{reg}} = \Sigma\left|\frac{X_t - F_t}{X_t}\right| / \, m(100) \tag{7}$$

Although the range of values that (7) can take is from 0 to $+\infty$, that of (6) is from 0 to 200%, or

$$0 \leq \text{MAPE}_{reg} \leq +\infty$$
$$0 \leq \text{MAPE}_{sym} \leq 200\%$$

Expression (6) provides a well defined range to judge the size of relative errors which may not be the case with expression (7). Expression (6) is also influenced by extreme values to a much lesser extent than expression (7).

1.5. The Median Absolute Percentage Error (MdAPE)

The Median Absolute Percentage Error is similar to MAPE (either regular or symmetric) but instead of summing up the Absolute Percentage Errors (APE) and then computing their average we find their median. That is, all the APE are sorted from the smallest to the largest and the APE in the middle (in case there is an even number of APEs then the average of the middle two is computed) is used to denote the median. The biggest advantage of the MdAPE is that it is not influenced by outliers. Its biggest disadvantage is that its meaning is less intuitive. An MdAPE of 8% does **not** mean that the average absolute percentage error is 8%. Instead it means that half of the absolute percentage errors are less than 8% and half are over 8%. (Using the symmetric APE reduces the chances of outliers and reduces the need to use MdAPE). Moreover, it is difficult to combine MdAPE across horizons and/or series and when new data becomes available.

1.6. Percentage Better (% Better)

The percentage better measure requires the use of two methods (A and B) and tells us the percentage of time that method A is better than method B (or vice versa). If more than two methods are to be compared the evaluation can be done for each pair of them. The range of % Better is from 0% to 100% (with 50% meaning a perfect tie between the two methods). As such it is an intuitive measure which provides precise information about the percentage of time that method A does better (or worse) than method B. The disadvantage of % Better is that it takes no account of the size of error assuming that small errors are of equal importance to large ones. In this respect it is not at all influenced by outliers. Its advantage, and value, comes, therefore, from cases when the size of errors is not important (e.g., in auctions) and when comparisons between **two** methods are desired.

1.7. The Average Ranking of Various Methods (RANKS)

Like the % Better measure RANKS requires at least two methods to compute. Similarly like % Better measure it ignores the size of errors. Instead the various methods are ordered (ranked) in inverse order to the size of their errors. Thus, the method with the smallest absolute error is given the value of 1, the one with the next smallest the value of 2 and so on, while the method with the largest absolute error is given the value m (where m is the total number of methods ranked). Consequently, the average of such RANKS is computed across methods and/or forecasting horizons. The biggest advantage of RANKS is that, like % Better, they are not influenced by extreme values. In addition they allow comparisons among any number of methods, where the % Better measure is limited to pairs of two only. Their biggest disadvantage is that their meaning is not intuitive. The average ranking can range from 1 to m. In the case that the average ranking is exactly (m + 1)/2 then all methods are similar. Methods whose ranking is less than (m + 1)/2 are doing better than average while those whose ranking is bigger are doing worse than average. However, it is not obvious through the RANKS how much better (or worse) a given method is in comparison to the others by simply examining the value of their RANKS. The biggest usefulness of RANKS is when the size of errors is not important, but picking the method which does most often better than the rest is (a piece of information useful in repeated auctions or any other case where the size of forecasting errors is of no importance).

1.8. Theil's U-Statistic (U-Statistic)

The Theil's U-Statistic (Theil, 1966) is defined as follows:

$$\mathrm{U-Statistic} = \sqrt{\sum_{t=1}^{m} \frac{\left(\frac{X_t - F_t}{X_t}\right)^2 / m}{\left(\frac{X_t - FN_t}{X_t}\right)^2 / m}} \tag{8}$$

where FN_t is some benchmark forecast such as the latest available value (the random walk, or Naive 1 forecast), or the latest available value after seasonality has been taken into account (Naive 2).

It can be noted that the numerator of (8) is similar to the numerator of expression (4), that is, the sum of percentage errors, e_t. However, as the percentage errors are square their absolute value is not needed since negative percentage errors will become positive once square. Similarly the denominator is the sum of percentage errors between the actual values and the benchmark forecasts.

Expression (8) simplifies to:

$$\text{U} - \text{Statistic} = \sqrt{\sum_{i=1}^{m} \frac{\left(\frac{X_t - F_t}{X_t}\right)^2}{\left(\frac{X_t - FN_t}{X_t}\right)^2}} \tag{9}$$

The range of expression (9) (or (8)) varies from 0 to ∞. A value of 1 means that the accuracy of the method being used is the same as that of the benchmark method. A value smaller than 1 means that the method is better than the benchmark while a value greater than one means the opposite.

The U-Statistic is greatly influenced by outliers. In the low end if $(X_t - F_t)/X_t$ is very small then its square is even smaller resulting in values very close to zero. On the upper end things are even worse. If FN_t is the same as X_t the denominator is zero, resulting in an infinite value of the U-Statistic when the numerator is divided by zero. Moreover, it is not obvious what a value of .85 means and how much better this value is than another one which is 0.82. Finally, although squaring the terms of expression (9) penalizes (like the RMSE) large errors, it can also result in outliers more often while making any interpretation of the U-Statistic less intuitive.

1.9. McLaughlin's Batting Average (Batting Average)

McLaughlin's (1975) Batting Average is an effort to make the U-Statistic more intuitive using two ways. First McLaughlin does not square the numerator and denominator of (9). Second, he defines the Batting Average as:

$$\text{Batting Average} = \left[4 - \sqrt{\sum_{i=1}^{m} \frac{\left|\frac{X_t - F_t}{X_t}\right|}{\left|\frac{X_t - FN_t}{X_t}\right|}}\right](100) \tag{10}$$

In which case 300 will mean similar performance as the benchmark, 300 to 400 better performance than the benchmark and less than 300 the opposite.

McLaughlin has attempted to make his Batting Average measure more intuitive by relating it to the batting average in baseball and by reducing the effect of outliers. However, when the actual value is very close to the benchmark forecast expression (10) can result in a negative value (if this happens it can be set to 400).

1.10. The Geometric Means of Square Error (GMMSE)

Geometric means average the product of square errors rather than their sums as in MSE. The geometric mean is therefore defined as

$$\text{GMMSE} = \left(\prod_t e_t^2\right)^{\frac{1}{m}} \tag{11}$$

Alternatively the Geometric Mean Root Mean Square Error (GMRMSE) can be found as follows:

$$\text{GMRMSE} = \left(\prod_t e_t^2\right)^{\frac{1}{2m}} \tag{12}$$

The biggest advantage of the geometric means is that the mean absolute errors of two methods (or models) can be compared by computing their geometric means. If one geometric mean is 10 and the other is 12 it can be inferred that the mean absolute errors of the second method are 20% higher than those of the first. In addition, geometric means are influenced to a much lesser extent from outliers than square means.

1.11. The Geometric Mean of Relative Absolute Errors (GMRAE)

The geometric mean of relative absolute errors is defined as

$$\text{GMRAE} = \left(\prod_t \text{RAE}_t\right)^{\frac{1}{m}} \tag{13}$$

where the RAE is computed as:

$$\mathrm{RAE}_t = \frac{\left|\frac{X_t - F_t}{X_t}\right|}{\left|\frac{X_t - FN_t}{X_t}\right|} \tag{14}$$

that is, the RAE is equivalent to the two terms of McLaughlin's Batting Average (see expression (10)). An alternative way of using (13) is by squaring the error terms of (14) in which case each RAE will be equivalent to Theil's U-Statistic (see expression (8)). The geometric mean root mean square error can also be found in a similar way to expression (12).

The advantage of the relative geometric means is that they are not contaminated as much by outliers and that they are easier to communicate than Theil's U-Statistic (Armstrong and Collopy, 1992). At the same time expression (14) is influenced by extremely low and large values. Armstrong and Collopy (1992) suggest Winsorizing the values of (14) by setting an upper limit of 10 and a low one of 0.01. Although the GMRAE might be easier to communicate than the U-Statistic it is still "typically inappropriate for managerial decision-making" (Armstrong and Collopy, 1992, p. 71).

1.12. Median Relative Absolute Error (MdRAE)

The median relative absolute error is found by ordering the RAE computed in (14) from the smallest to the largest and using their middle value (the average of the middle two values if m is an even number) as the median. In this respect the MdRAE is similar to the MdAPE except that expression (14) is used to compute the error used in finding the median rather than the APE.

The advantage of the MdRAE is that it is not influenced by outliers while allowing comparisons with a benchmark method. Its disadvantage, as that of the MdAPE, is that its meaning is not clear -- even more so than that of MdAPE.

1.13. Differences of APE of Naive 2 Less APE of a Certain Method (dMAPE)

The difference in the Absolute Percentage Error (APE) of Naive 2 (deseasonalized random walk) minus the APE of a certain method can be computed as:

$$\mathrm{dMAPE} = \frac{\Sigma\left[\left|\frac{X_t - FN_t}{X_t}\right| - \left|\frac{X_t - F_t}{X_t}\right|\right]}{m} \tag{15}$$

or better the differences in the symmetric MAPE can be found as:

$$\mathrm{dMAPE}_{sym} = \frac{\Sigma\left[\left|\frac{X_t - FN_t}{(X_t + FN_t)/2}\right| - \left|\frac{X_t - F_t}{(X_t + F_t)/2}\right|\right]}{m} \tag{16}$$

The dMAPE tells us how much better (in absolute percentage terms) or worse the forecasts of some methods are than those of Naive 2 (or Naive 1, i.e., random walk) or some other method. The dMAPE measure is relative and intuitive (negative values mean that the method does worse than Naive 2, positive values mean better). Furthermore, there is practically never the chance of dividing by zero, as it is the case with GMRAE, MdRAE, U-Statistic or Batting Average.

1.14. R^2

R^2 is used a great deal in regression analysis and is defined as the ratio of the explained to the total variation, or

$$R^2 = \frac{\Sigma EE_t^2}{\Sigma TE_t^2} \tag{17}$$

where

EE_t is the explained error at t and is defined as $F_t - \overline{X}_t$ (i.e., the difference of the forecast minus the mean of the X values), and

TE_t is the total error at t, or $X_t - \overline{X}_t$ (i.e., the actual value minus the mean).

In this respect R^2 refers to forecasting errors in relation to a benchmark, the mean. R^2 fluctuates between 0 and 1 and since it is the outcome of a ratio it tells us the percentage of the total variation (errors square) explained by the forecasting method in relation to the mean.

The biggest advantage of R^2 is that it is a relative measure that is easy and intuitive to understand. Its disadvantage is that the benchmark is the mean which makes it inappropriate when there is a strong trend in the data necessitating alternatives, like Theil's U-Statistic, which are more appropriate for data with a strong trend. R^2 is used a great deal in regression analysis but has found no place in forecasting. It will not, therefore, be used in this study.

1.15. Classifying the Various Methods

Table 1 classifies the fourteen methods discussed above according to two criteria (the character of the measure and the type of evaluation).

Table 1: Classifying the Major Accuracy (Error) Measures

		Evaluation is Done		
		On a Single Method	On More than One Method	In Comparison to Some Benchmark
Character of Measure	Absolute	MSE MAE GMMSE	RANKS	
	Relative to a Base or other Method		% Better	U-Statistic Batting Average GMRAE MdRAE
	Relative to the Size of Errors	MAPE MdAPE	dMAPE MAPE MdAPE	R^2 dMAPE

It is important to note that all of the fourteen accuracy measures included in Table 1 are unique either in the loss function they use, or their character/type of evaluation. Each provides, therefore, some distinctive information/value that needs to be traded off against possible disadvantages.

2. EVALUATING THE VARIOUS ACCURACY MEASURES

Each of the fourteen accuracy measures discussed in the last section provides us with some unique information. It can be, therefore, argued that they are all, in some way, useful and that they should all be used collectively. At the same time it is practically impossible to use fourteen measures. We must, therefore, develop criteria for their evaluation. In this study we are using two statistical and two user related criteria. The statistical criteria refer to the reliability and discrimination of a measure while the non-statistical ones examine their information content and intuitiveness. However, as we will demonstrate it is not possible to optimize these criteria at the same time, requiring us to consider the tradeoffs involved.

From a statistical point of view a measure must be reliable and able to discriminate appropriate models or methods from inappropriate ones; although it is possible that reliable measures may not be discriminating enough and vice versa.

2.1. Statistical Criteria

Statistical measures need to be both reliable and discriminating. Reliability is defined as the ability of a measure to produce as similar results as possible when applied to different subsamples of the same series. In such a case variations in the accuracy measure are the result of differences in the series contained in each subsample and can be referred to as "Within (series) Variation". The **smaller** the within variation the better, as it implies that the measure used is not influenced by the specific series contained in each subsample, by extreme values, or other characteristics of the individual series. A measure is consistent (reliable) when it is not much influenced by within series fluctuations (and vice versa). At the same time accuracy measures should be capable of discriminating between appropriate and less appropriate models or methods. This means that if different methods (or models) are used with the **same** set of series then the most discriminating measure will be the one that produces the **highest** variations in the accuracy between these methods. Thus, the larger the "Between (methods) Variation" the better as it implies that the measure used is capable of discriminating among methods (or models) by telling us which is the most appropriate among them.

Ideally we would prefer accuracy (or error) measures which are as reliable as possible **and** as discriminating as possible, although this may not always be an attainable objective. For instance, Figure 1 shows the values of the MdRAE for the method of Single exponential smoothing when the 1001 series of the M-Competition (Makridakis et al., 1982) have been subdivided into nine subsamples of 111 series each. Figure 2 shows similar values but this time using MSE. Obviously the reliability of MdRAE is practically perfect as all nine subsamples provide practically the same values for this method. There are **no** fluctuations in the MdRAE values until the eighth forecasting horizon and then the MdRAE becomes 0.99 for horizons 9 and 16, and 1.02 for horizon 18. At the other extreme, the values of MSE vary widely making the MSE an unreliable measure, indicating that it is greatly influenced by some of the series contained in each subsample. For example, the MSE of subsample 5 are about four times as big as those of the other subsamples whose values also fluctuate a great deal.

Figure 3 shows the MdRAE when nine different methods are used to estimate each of the 1001 series while Figure 4 shows the same information using MSE. The fluctuations in the MdRAE are again much smaller than those of MSE. If the MdRAE of Regression are excluded the range of the remaining ones fluctuates little. Although the MdRAE measure is highly reliable, it does not discriminate enough to confidently tell us which method(s) is(are) better than others. The MSE, on the other hand, can better discriminate among methods as the values shown in Figure 4 vary considerably from one method to another (the MSE scale in Figure 4 is in thousands). It follows that if the only two accuracy measures available were the MdRAE and the MSE, then it would have been impossible to say which one of them was the most appropriate from a statistical point of view. We need, therefore, to determine a way to compare the reliability and discrimination of these various measures.

2.1.1. *Within and Between Coefficients of Variation (C of V)*

Table 2(a) shows the MdRAE for each of the nine subsamples using the method of Single exponential smoothing. In addition it shows the overall mean, standard deviation, and coefficient of variation for these nine subsamples. Table 2(b), on the other hand, shows the MdRAE for the nine different methods used in this study, together with the overall mean, standard deviation and coefficient of variation. The "Within" method's coefficient of variation of Table 2(a) tells us how much each of the nine subsamples

FIGURE 1
MdRAE: WITHIN 9 SUBSAMPLE(SS) OF SINGLE SMOOTHING

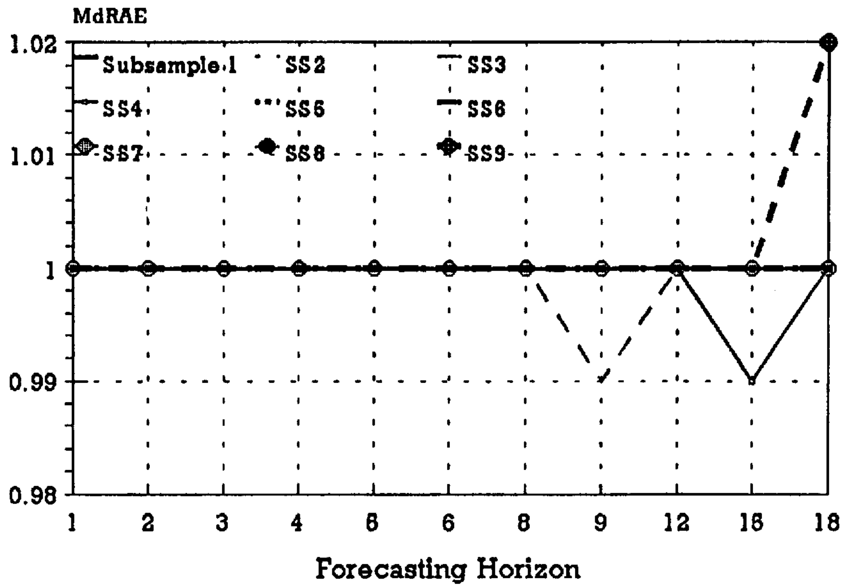


FIGURE 2
MSE: EACH OF 9 SUBSAMPLE(SS) OF SINGLE SMOOTHING

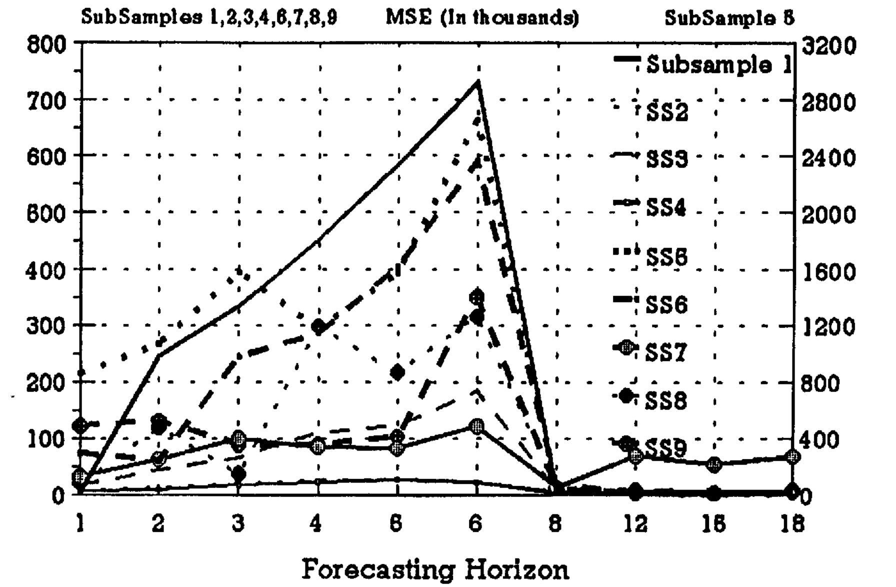


FIGURE 3
MdRAE (ALL SERIES): BETWEEN METHODS

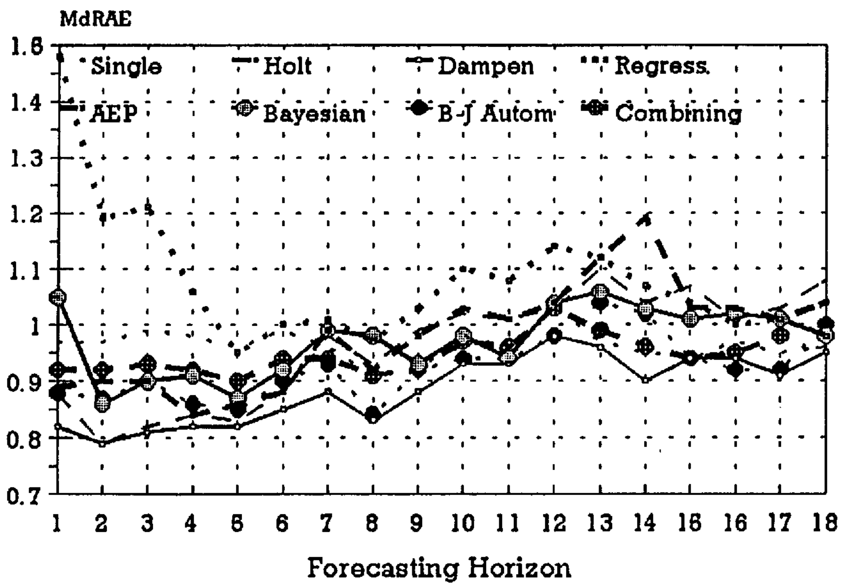


FIGURE 4
MSE (ALL SERIES): BETWEEN METHODS

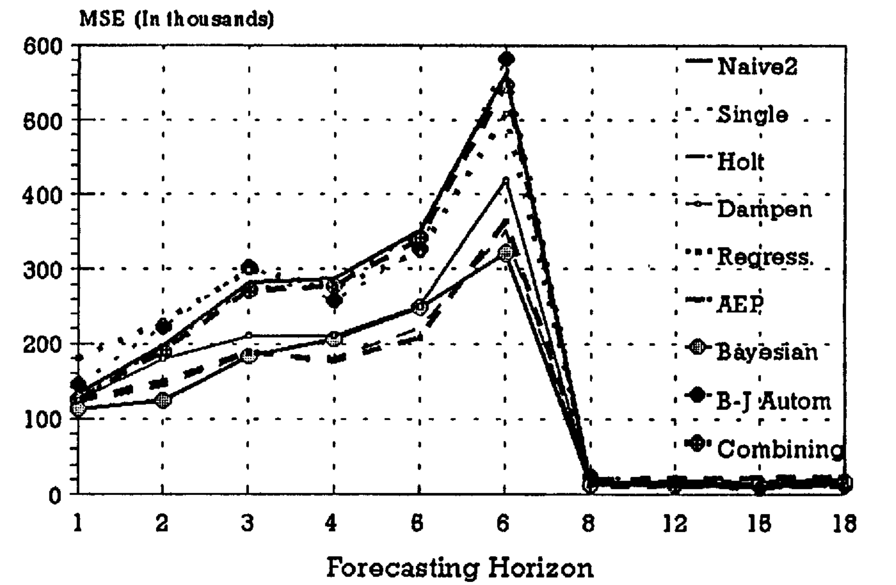


Table 2(a)

METHODS	FORECASTING HORIZONS																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Subsample 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Subsample 2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Subsample 3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Subsample 4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
Subsample 5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Subsample 6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Subsample 7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Subsample 8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Subsample 9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.02	1.00	1.00	1.00	1.00	1.00	1.00
Average	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Standard Dev.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Coefficient of Variation	0.03	0.00	0..00	0.00	0.03	0.00	0.03	0.00	0.42	0.00	0.00	0.52	0.09	0.00	0.19	0.03	0.07	0.09

Table 2(b)

F/C H/n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Average
AEP	0.89	0.9	0.9	0.84	0.86	0.88	0.99	0.92	0.98	1.03	1.01	1.04	1.12	1.19	1.03	1.03	1.01	1.04	0.963
Bayesian	1.05	0.86	0.9	0.91	0.87	0.92	0.99	0.98	0.93	0.98	0.94	1.04	1.06	1.03	1.01	1.02	1.01	0.98	0.963
Combining	0.92	0.92	0.93	0.92	0.9	0.94	0.94	0.91	0.92	0.97	0.96	1.03	0.99	0.96	0.94	0.95	0.98	1	0.944
Damped	0.82	0.79	0.81	0.82	0.82	0.85	0.88	0.83	0.88	0.93	0.93	0.98	0.96	0.9	0.94	0.94	0.91	0.95	0.873
Holt	0.88	0.79	0.82	0.84	0.83	0.88	0.98	0.93	0.99	1.02	1.01	1.03	1.1	1.04	1.07	1.01	1.03	1.08	0.942
B-J	0.88	0.87	0.9	0.86	0.85	0.9	0.93	0.84	0.92	0.94	0.94	0.98	1.04	1.02	0.94	0.92	0.92	1	0.916
Regression	1.48	1.19	1.21	1.06	0.95	1	1.01	0.97	1.03	1.1	1.08	1.14	1.12	1.07	1.04	1	1.01	0.98	1.089
Single	0.97	0.97	0.99	0.98	0.96	0.99	0.96	0.91	0.92	0.98	0.96	1.03	0.97	0.95	0.94	0.97	0.95	0.96	0.966
Average	0.986	0.911	0.932	0.904	0.880	0.920	0.960	0.911	0.946	0.994	0.979	1.034	1.045	1.020	0.989	0.980	0.978	0.999	0.957
Stand Dev	0.197	0.120	0.118	0.077	0.049	0.050	0.039	0.050	0.046	0.051	0.048	0.046	0.062	0.083	0.051	0.038	0.043	0.040	0.084
Coef of V	20.02	13.16	12.64	8.53	5.57	5.46	4.10	5.53	4.84	5.18	4.88	4.48	5.90	8.11	5.17	3.89	4.36	4.02	7.579

fluctuates around the overall mean of all the 1001 series, while the "Between" method's coefficient of variation of Table 2(b) tells us, correspondingly, how much each of the nine methods fluctuates around the overall mean of these nine methods. The smaller the values of the "Within" coefficient of variation the more reliable is the measure of MdRAE for Single smoothing, the bigger the "Between" coefficient of variation the more it can discriminate in helping us identify the more appropriate of the method(s) (or models) being compared. As the coefficients of variation are relative measures they allow us to compare the "Within" and "Between" variations of MdRAE for different methods as well as those of MdRAE with the other measures whose means and standard deviations are unequal. They also permit us to average across methods and horizons.

Figure 5 shows a graph of the coefficients of variation listed in Table 2(a) and 2(b). Its biggest advantage is that it allows us to visualize (study) the criteria of both reliability **and** discrimination at the same time. Comparing Figure 5 to those of 6 to 8 provides us with different information concerning the statistical properties of the MdRAE, % Better, RANKS and $MAPE_{sym}$ for Single exponential smoothing. The "Within" and "Between" fluctuations of these four methods vary considerably suggesting the need to consider trading off reliability for discrimination and vice versa. For instance the coefficient of variation for $MAPE_{sym}$ (Figure 8) is about twice as much as that of RANKS while the within variation of % Better becomes proportionally larger for longer forecasting horizons while it is less than the "Between" for the first four ones. Figures similar to those of 5 to 8 can be made for all accuracy measures and methods. As the number of graphs involved approaches 100 they cannot be present. However, those of Figures 5 to 8 present typical cases of "Between" and "Within" variations.

Within vs Between Coefficients of Variation: The coefficients of variation shown in Figures 5 to 8 can be displayed by plotting the "Between" versus the "Within" variations in a way that allows us to directly see the relationship (and tradeoffs) between the two. Notice that the scale of the "Between" variation goes from a high value to zero. This type of plot facilitates considering the tradeoff by wanting values along the diagonal line as close to the origin of the XY axis as possible. Figure 9, for instance, shows that the "Within" variation of Single smoothing does not vary while the "Between" one decreases with longer forecasting horizons. The first forecasting horizon of % Better (Figure 10) shows high "Within".

FIGURE 5
MdRAE: WITHIN AND BETWEEN VARIATION
SINGLE SMOOTHING: 1-18 F/C HORIZONS

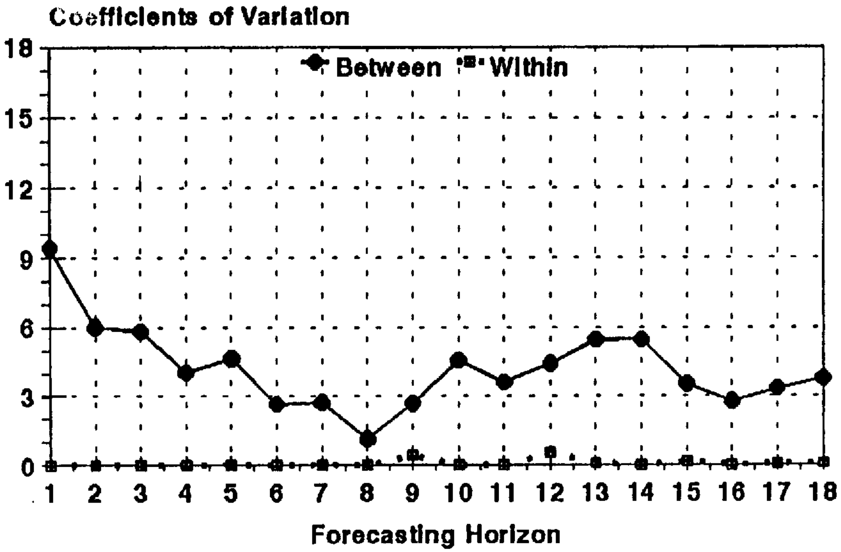


FIGURE 6
% BETTTER: WITHIN AND BETWEEN VARIATION
SINGLE SMOOTHING: 1-18 F/C HORIZONS

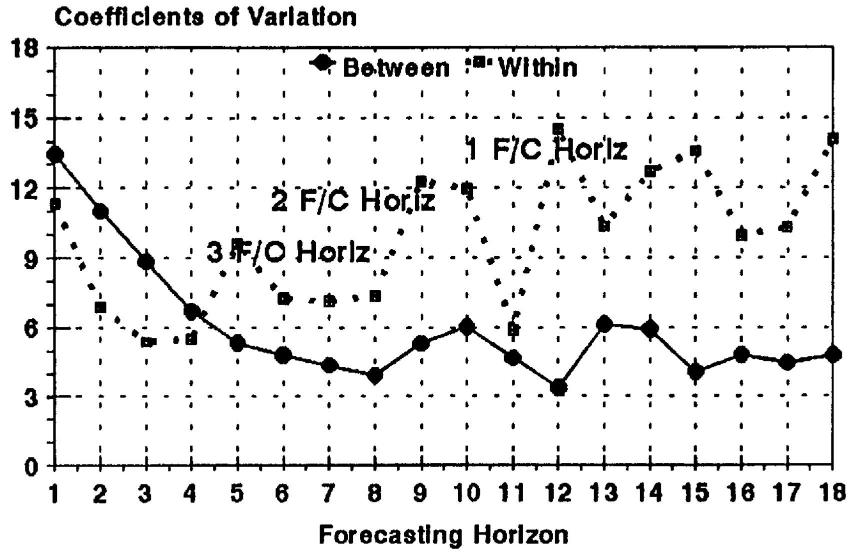


FIGURE 7
RANKS: WITHIN AND BETWEEN VARIATION
SINGLE SMOOTHING: 1-18 F/C HORIZONS

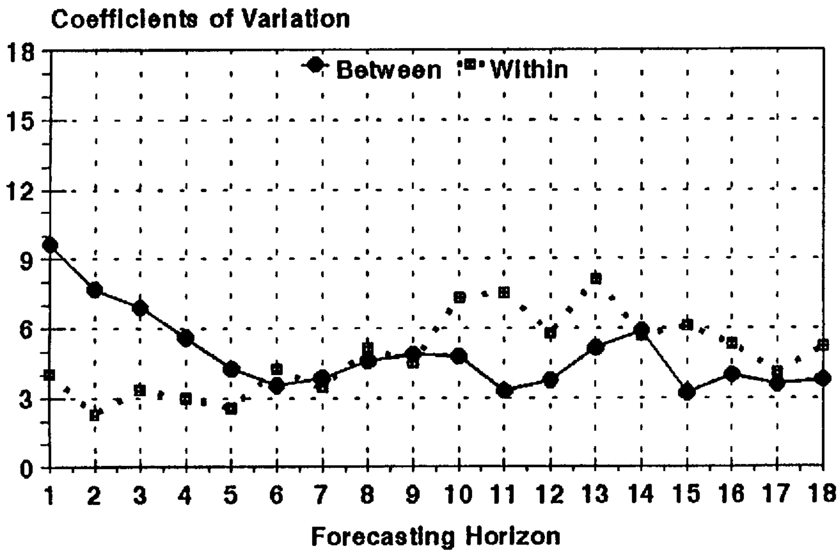


FIGURE 8
MAPEsym: WITHIN AND BETWEEN VARIATION
SINGLE SMOOTHING: 1-18 F/C HORIZONS

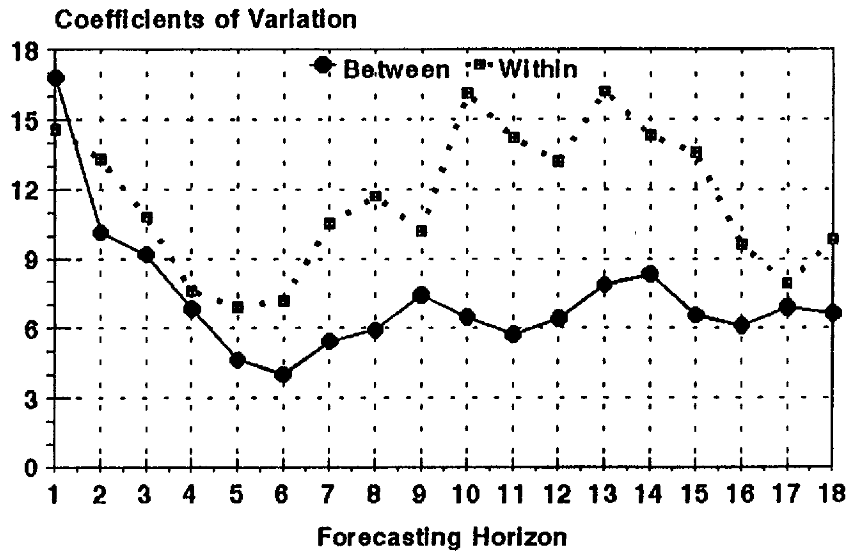


FIGURE 9
MdRAE: WITHIN AND BETWEEN VARIATION
SINGLE SMOOTHING

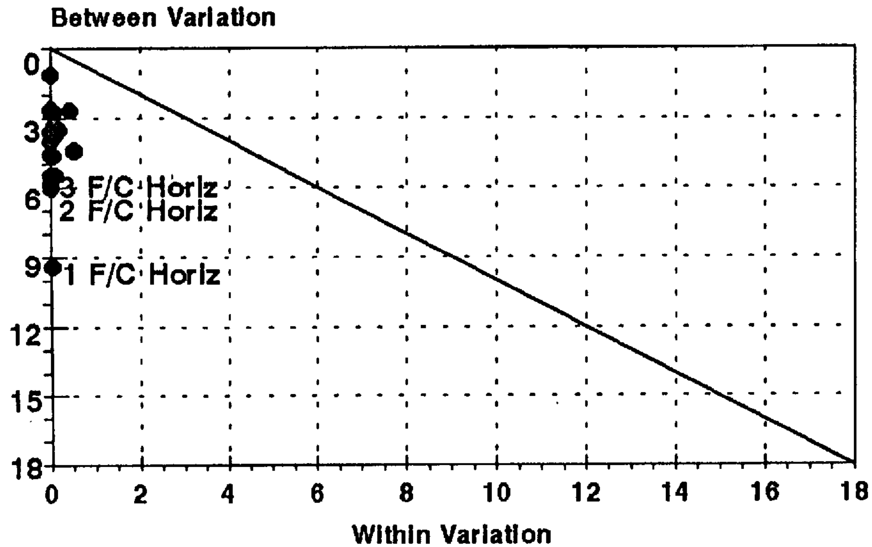


FIGURE 10
% BETTTER: WITHIN AND BETWEEN VARIATION
SINGLE SMOOTHING

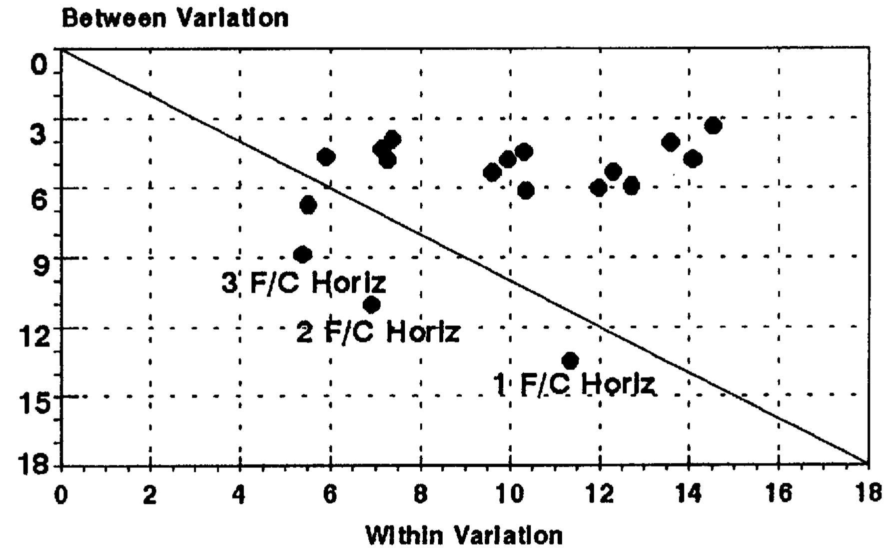


FIGURE 11
RANKS: WITHIN AND BETWEEN VARIATION
SINGLE SMOOTHING

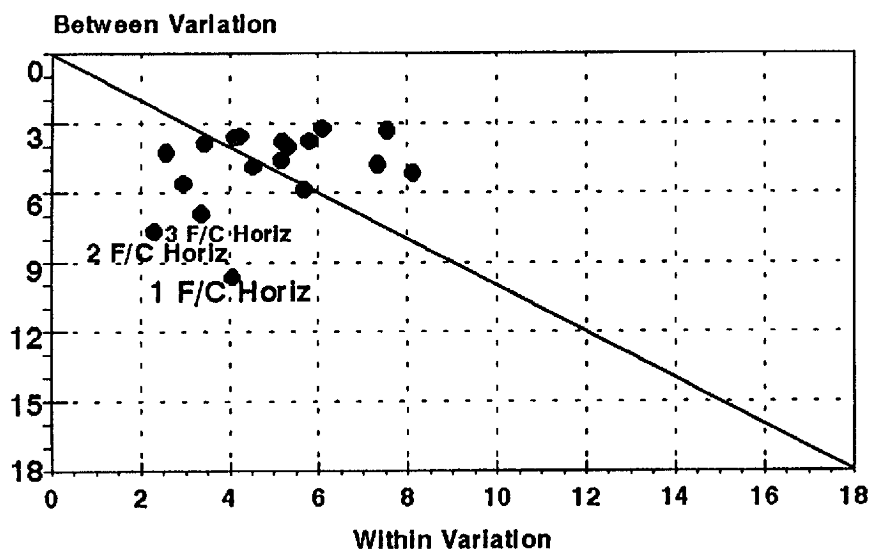
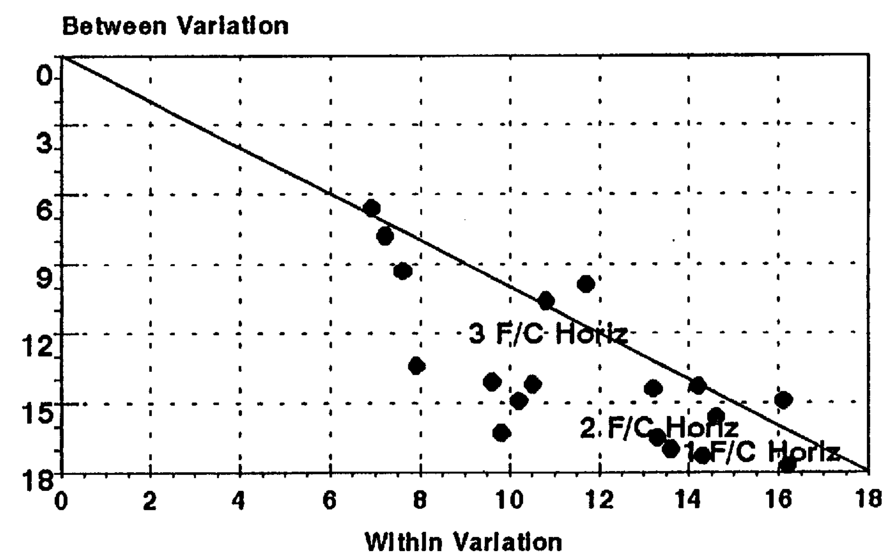


FIGURE 12
MAPEsym: WITHIN AND BETWEEN VARIATION
SINGLE SMOOTHING



Consequently, for horizons 2 and 3 not only the "Within" but also the "Between" decreases. However, for longer forecasting horizons "Within" increase without a corresponding decrease in "Between". This is not obviously a desired characteristic as it means a decrease in reliability without a corresponding improvement in discrimination. The same is true, but to a lesser extent, with RANKS (Figure 11) while the variations of $MAPE_{sym}$ (Figure 12) indicate that as one increases the other decreases (on average) and vice-versa. This type of behavior of $MAPE_{sym}$ makes the tradeoffs between reliability and discrimination very clear.

Averaging Across Methods: The "Within" and "Between" variations of each measure can be averaged across the nine methods used in this study. This will result in a single graph for each of the thirteen accuracy measures being considered. That is, graphs similar to that of Figure 9 (MdRAE for Single exponential smoothing) can be constructed for all nine methods for, say, the measure of MdRAE (see Figure 13). Consequently, graphs can also be made for all thirteen accuracy measures studied in this paper. Figures 14, 15 and 16 show the average across methods of the "Within" and "Between" variations for the measures of % Better, RANKS and $MAPE_{sym}$.

Averaging Across Forecasting Horizons: The "Within" and "Between" variations of each measure can be averaged across the various forecasting horizons (six for yearly, eight for quarterly and eighteen for monthly data) used. This average will result in the "Within" and "Between" variations for each of the nine methods being used for each accuracy measure. Figures 17, 18, 19 and 20 show such averages for the accuracy measures of MdRAE, % Better, RANKS and $MAPE_{sym}$. Apart from MdRAE there is not much fluctuation in the coefficients of variation regarding the nine methods. That is the within variation of a single accuracy measure varies very little from one method to another with the only exception of MdRAE. At the same time Figures 17-20 indicate that some measures are more reliable for some methods than others. For example RANKS provide more reliable results for Holt's smoothing than $MAPE_{sym}$. On the other hand, AEP discriminates better than Box-Jenkins or Regression.

FIGURE 13
MdRAE: WITHIN AND BETWEEN VARIATION
AVERAGE ALL METHODS

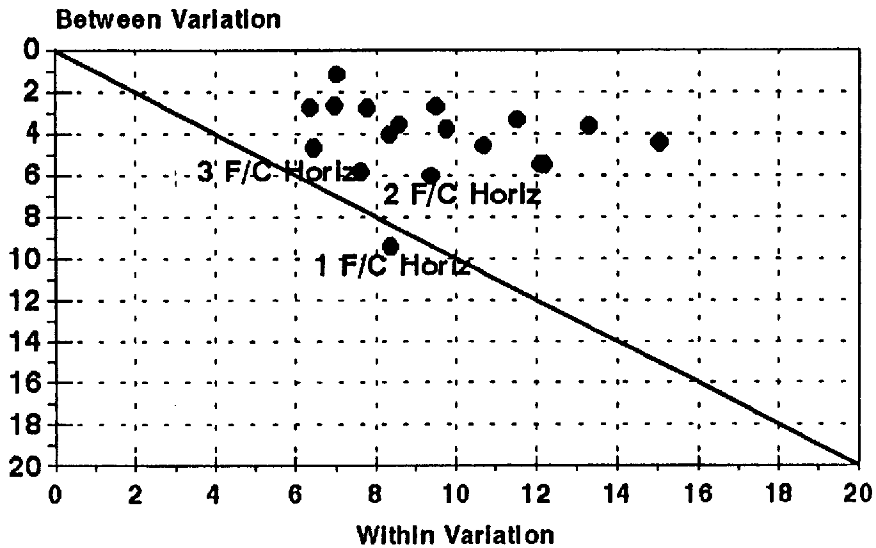


FIGURE 14
% BETTER: WITHIN AND BETWEEN VARIATION
AVERAGE ALL METHODS

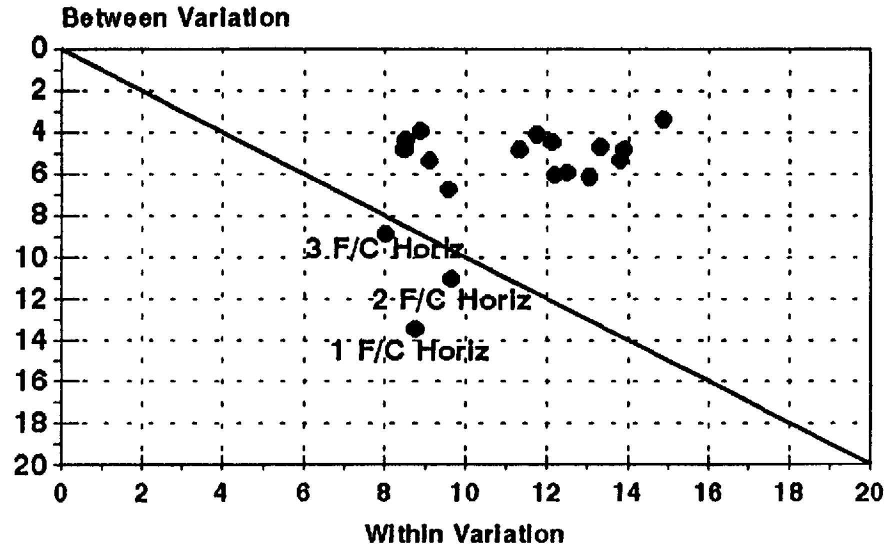


FIGURE 15
RANKS: WITHIN AND BETWEEN VARIATION
AVERAGE ALL METHODS

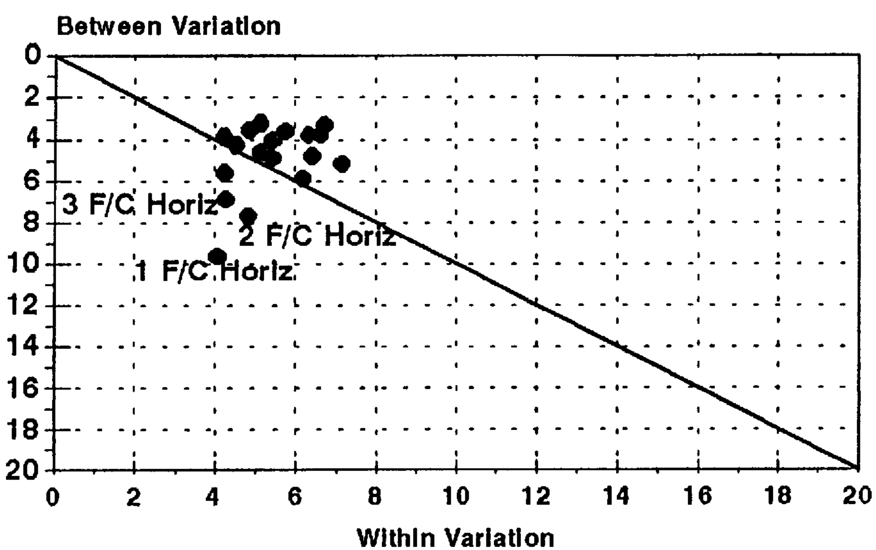


FIGURE 16
MAPEsym: WITHIN AND BETWEEN VARIATION
AVERAGE ALL METHODS

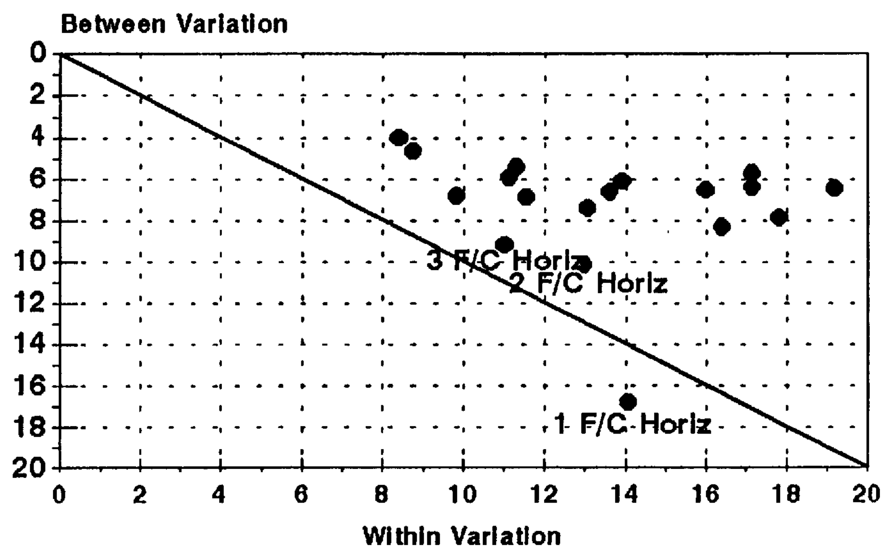


FIGURE 17
MdRAE: WITHIN AND BETWEEN VARIATION
AVERAGE ALL FORECASTING HORIZONS

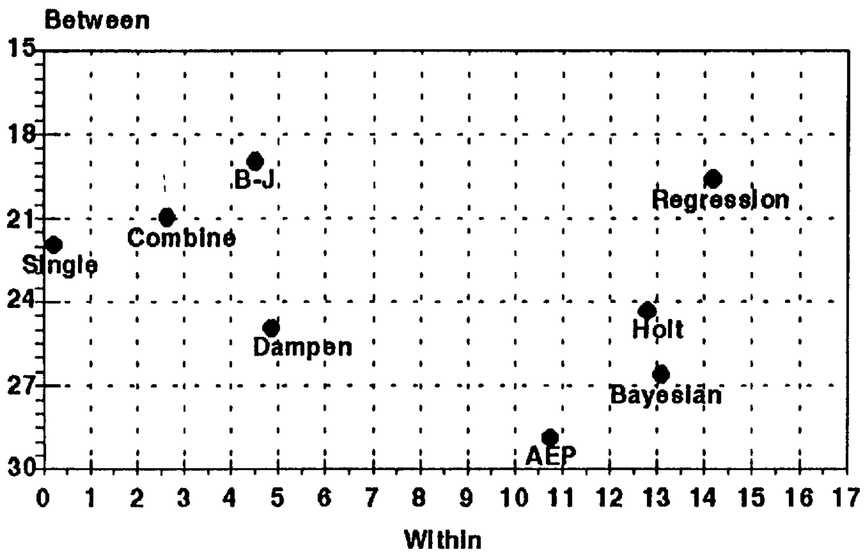


FIGURE 18
%BETTER: WITHIN AND BETWEEN VARIATION
AVERAGE ALL FORECASTING HORIZONS

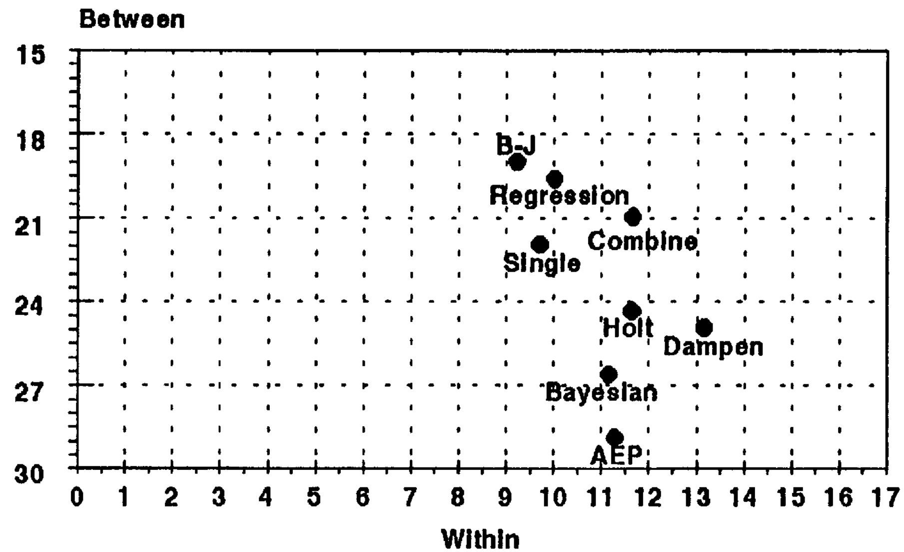


FIGURE 19
RANKS: WITHIN AND BETWEEN VARIATION
AVERAGE ALL FORECASTING HORIZONS

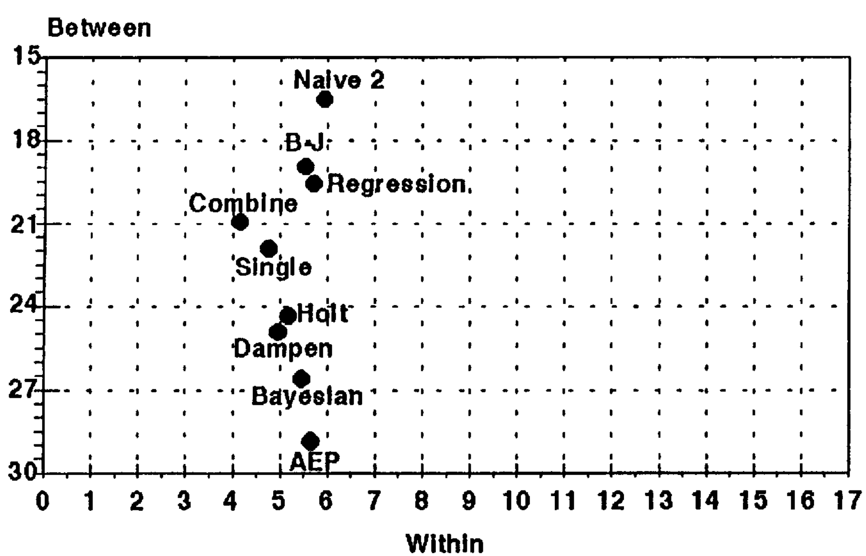
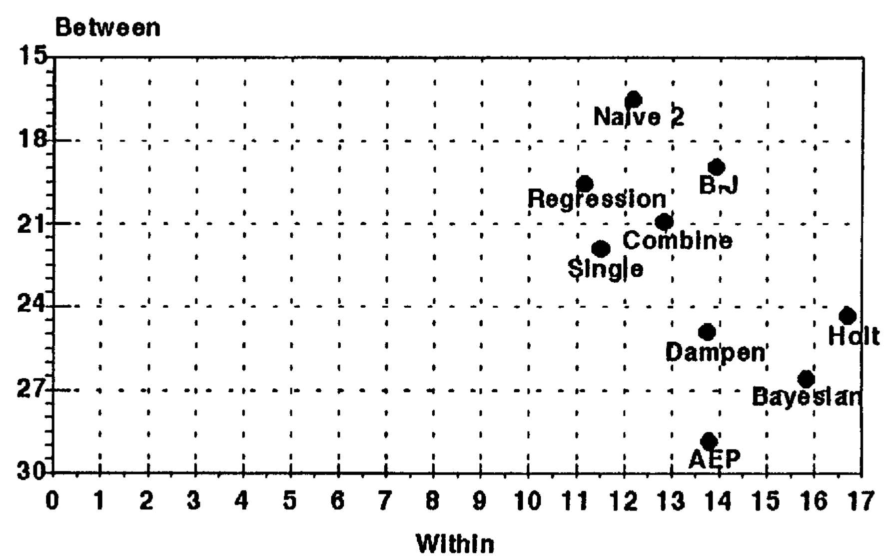


FIGURE 20
MAPEsym: WITHIN AND BETWEEN VARIATION
AVERAGE ALL FORECASTING HORIZONS



Average Across Methods and Forecasting Horizons: Averaging across methods **and** forecasting horizons produces a single "Within" and a single "Between" value for each of the thirteen accuracy measures. These variations are shown in Table 3.

Table 3

Accuracy Measure	Variation	
	Within	**Between**
1. Mean Square Error (MSE)	161.2	19.8
2. Mean Absolute Error (MAE)	137.7	18.0
3. Regular Mean Absolute Percentage Error ($MAPE_{reg}$)	190.5	98.8
4. Symmetric Mean Absolute Percentage Error ($MAPE_{sym}$)	13.3	7.3
5. Median of the Absolute Percentage Error (MdAPE)	17.2	7.8
6. % of Times Method A is Better than Method B (% Better)	10.8	6.6
7. Average RANKS (RANKS)	5.3	5.3
8. Theil's U-Statistic (Theil's-U)	6.0	5.3
9. McLaughlin's Batting Average (Batting Average)	1.6	1.6
10. Geometric Mean Square Error (GMMSE)	15.4	5.1
11. Geometric Mean Absolute Relative Error (GMRAE)	12.0	8.2
12. Median of the Relative Absolute Errors (MdRAE)	9.3	4.6
13. Difference of the symmetric MAPE of Naive 2 minus some Method ($dMAPE_{sym}$)	10.8	9.7
Average	33.4	8.3

Figure 21 shows a graph of the values listed in Table 3, while Figure 22 shows all measures except for those of MSE, MAE and $MAPE_{reg}$ which, as can be seen in Figure 21, are considerably bigger than the rest. Figures 21 and 22 tell us a great deal about the statistical properties of reliability and discrimination of the thirteen measures (for the average of all methods and forecasting horizons) included in this study. Clearly the $MAPE_{reg}$, the MAE and the MSE are in a category of their own. The regular MAPE is the most discriminating and the least reliable measure followed by the MSE and the MAE. If the only accuracy measures available were the $MAPE_{reg}$, the MAE and the MSE, it is not obvious which of the three is the most appropriate for both selecting a model and comparing various methods as there are trade offs to be made between reliability and discrimination.

When the remaining accuracy measures are seen (Figure 22) without the $MAPE_{reg}$, MSE and MAE it is clear that some measures are suboptimal, from a statistical point of view, as they exhibit more "Between" fluctuations without greater "Between" (or vice versa) than some other measures. The measures in the efficiency frontier are denoted with a circle and are connected with a dotted line while the suboptimal

FIGURE 21
WITHIN AND BETWEEN VARIATION: AVERAGE ALL METHODS AND FORECASTING HORIZONS

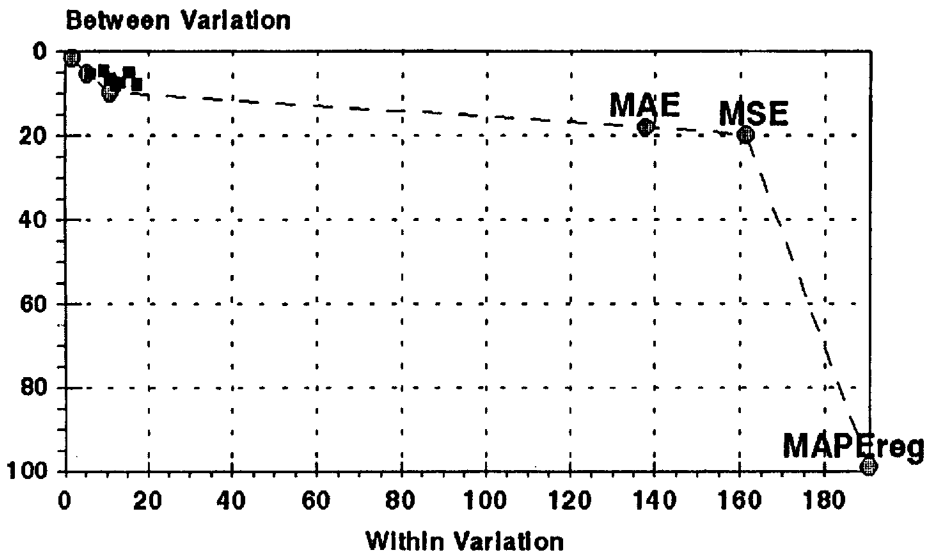


FIGURE 22
WITHIN AND BETWEEN VARIATION: AVERAGE ALL METHODS AND FORECASTING HORIZONS (NO MAPEreg, MSE AND MAE)

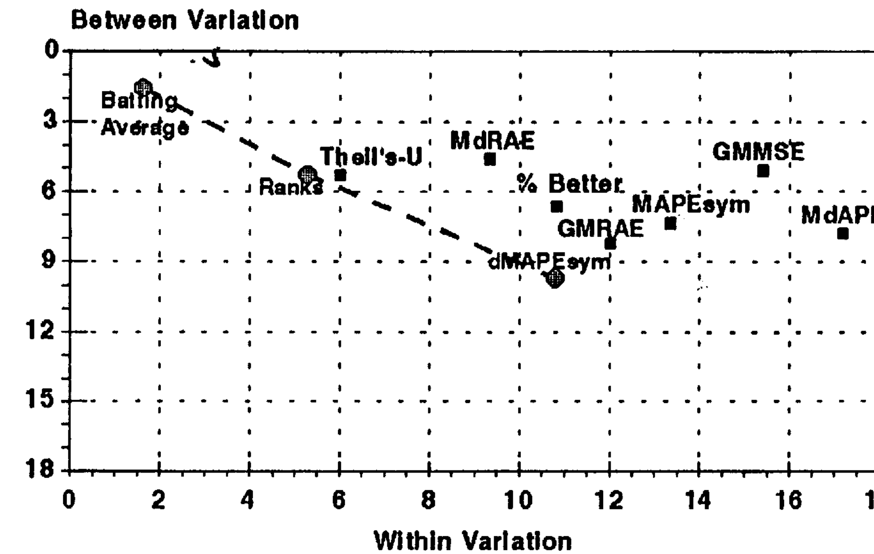


FIGURE 23
WITHIN AND BETWEEN VARIATION: ALL METHODS/HORIZONS BAYESIAN FORECASTING EXCLUDED

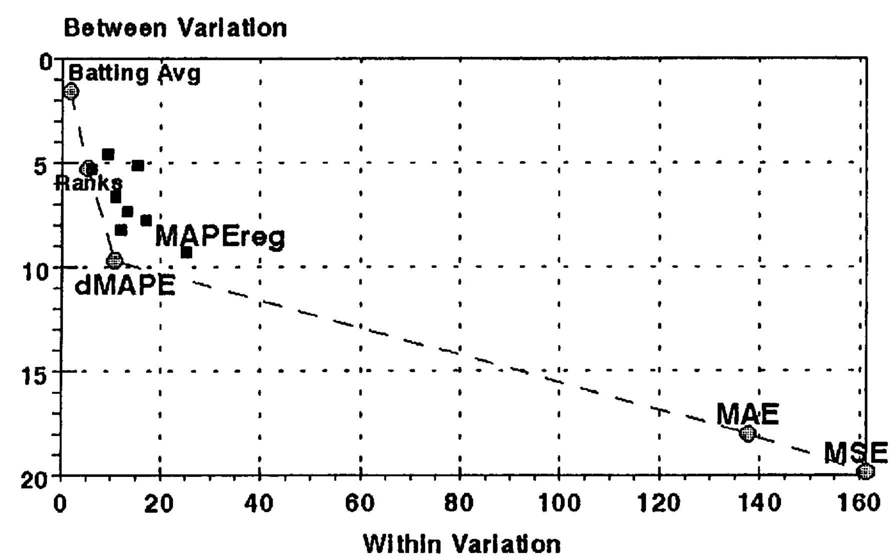
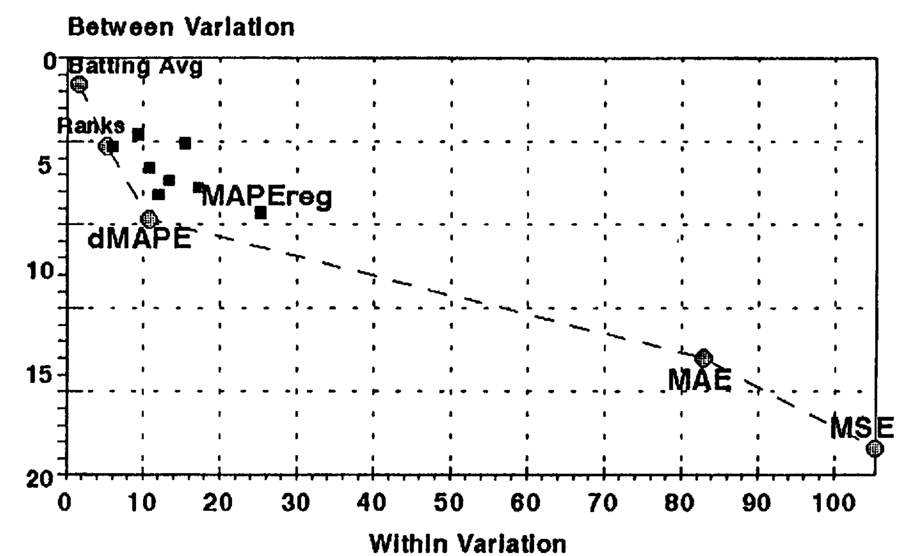


FIGURE 24
WITHIN AND BETWEEN VARIATION: ALL METHODS/HORIZONS SIX OUTLIERS ELIMINATED FROM MAPEreg AND MAE/MSE



ones are shown with a small square. Unless there are non-statistical reasons it makes little sense to utilize the suboptimal accuracy measures for making comparisons among methods and/or selecting the most appropriate model.

2.1.2. *The Effect of Outliers*

Outliers can influence a great deal the $MAPE_{reg}$, the MSE and the MAE. In the case of $MAPE_{reg}$ five series in a single method (Bayesian forecasting) are responsible for the extremely high value of both the within and between variation of $MAPE_{reg}$. If these five extremely large percentage errors (caused by a very small X_t value coupled with a large e_t value) of Bayesian forecasting are excluded both the within and between variation of $MAPE_{reg}$ are reduced substantially (see Figure 23) bringing $MAPE_{reg}$ close to the remaining measures.

Both the MSE and MAE are also influenced by outliers. This time, however, the outliers are six series which affect equally practically all methods (five out of these six series caused extremely large errors in all nine methods). Eliminating these six series reduces the within variation of MSE and MAE (see Figure 24) but affects much less the between methods variation as all methods are influenced in practically the same manner by the same six series. By eliminating more outliers the within variation of MSE and MAE is further reduced but by relatively small amounts, leaving always the MSE and MAE in a category of their own -- much higher than the remaining eleven measures.

Outliers, as has been shown, influence some accuracy measures much more than others resulting in values which are not representative of the typical errors and which can lead to inappropriate decisions. For these reasons outliers, or their effect must be eliminated. To better appreciate the importance of outliers we consider the effect of a single outlier when 50 values or errors are being used (see Table 4). MSE is influenced the most while the GMMSE, another quadratic loss function, is influenced the least.

Table 4
The Effect of Large and Small Errors on the Accuracy Measures of MSE, MAE, GMMSE, MAPE_{reg} and MAPE_{sym}

Size of Errors		MSE		GMMSE		MAPE_{reg}		MAPE_{sym}	
Actual Size	No. of Standard Deviations	Actual Size	% Increase from Base	Actual Size	% Increase from Base	Actual Size	% Increase from Base	Actual Size	% Increase from Base
10.1	½	1841	-.3%	997	-2.6%	4.00%	-6.7%	3.98%	-5.6%
20.2	1	1847	Base	1025	Base	4.29%	Base	4.22%	Base
40.3	2	1871	1.3%	1054	2.8%	4.86%	13.4%	4.61%	9.3%
60.5	3	1913	3.6%	1071	4.6%	5.44%	26.9%	4.92%	16.7%
100.8	5	2043	10.6%	1093	6.7%	6.59%	53.8%	5.39%	27.8%
201.7	10	2655	43.7%	1124	9.8%	9.47%	121.0%	6.07%	44%
302.5	15	3694	100.0%	1142	11.5%	12.35%	188.3%	6.45%	52.91%

Table 4 indicates that something needs to be done to protect against extreme values in the MSE. At the same time it must be understood that methods which are not much influenced by extreme values cannot easily recognize such values and therefore discriminate enough when they are present so that an appropriate model or method can be selected. Again there is a tradeoff between extreme value(s) negatively influencing certain measures and the ability of such measures to discriminate among various models/methods.

Apart from MAPE_{reg}, MSE and MAE outliers do not greatly affect the remaining accuracy measures apart from the following special cases:

1. When $X_t - FN_t = 0$, the measures of Theil's-U, Batting Average and GMRAE do not work as there is a divisor by zero (see expressions (9), (10) and (14)). Similarly if $X_t - FN_t$ is very small (9), (10) and (14) can result in large values causing outliers.

2. When $X_t - F_t = 0$, the GMMSE and the GMRAE cannot be computed as the product of (11) or (13) becomes zero.

3. When $X_t = 0$, the $MAPE_{reg}$, Batting Average and Theil's U-Statistic cannot be computed as there is a division by zero.

4. When F_t is much bigger than X_t the $MAPE_{reg}$ can be greatly affected, causing outliers.

For the measures of Theil's U, Batting Average and GMRAE, it has been suggested to winsorize their values by setting lower and upper limits.

For many measures when $X_t = 0$ its value can be set to an arbitrary small value whereas when F_t is much bigger than X_t the $MAPE_{sym}$ ought to be used.

There are no easy solutions to avoiding the negative consequences of outliers as far as MSE is concerned apart from using the GMMSE which is not influenced very much by extreme errors. On the other hand, the methods of MdAPE and MdRAE are not influenced by outliers while the methods of $dMAPE_{sym}$ still allows comparisons with a benchmark forecast without, however, the possibility of a zero divisor.

2.2 User Oriented Criteria

Although statistical considerations are critical for judging the value of the various accuracy measures, they do not suffice. For these measures to be useful they must also be understood and used -- most often by people with little or no statistical background. Accuracy measures need, therefore, to provide useful information to decision makers and, at the same time, be intuitive.

2.2.1. *Informativeness*

All statistical measures we examined in this paper are unique (with the exception of $MAPE_{reg}$ and $MAPE_{sym}$ which are actually a single measure being computed in two different ways) in the information they provide. This can be seen in Table 1, classifying these measures, where even when there is more than one measure in a single box the information provided by them is not the same because the loss function used to compute them is different. For instance MSE computes the average of the square errors (i.e., it employs a quadratic loss function) while the MAE calculates the average of the absolute errors. Similarly, the MSE differs from the GMMSE, although both use quadratic losses, as the latter computes the average of the product of errors rather than the average of the sum of such errors. This uniqueness of accuracy measures justifies Winkler and Murphy's (1992) observation that no measure can be excluded a priori. At the same time we know that some accuracy measures have passed the acid test of time and they are used to a much greater extent than others. This is not by chance but rather because they provide some specific and useful information, to decision or policy makers, not available by the other measures.

We believe/know that the two measures used more than any others are the MAPE and the MSE. The former is used to report results **and** make relative comparisons among various methods and/or periods/situations. The latter is used as input to inventory and material requirement systems to take into account the uncertainty in the forecasts.

The informativeness of accuracy measures relates to their intended use as some aim at reporting or using the results of forecasting methods while others at making comparisons among methods -- including evaluating the results of forecasting competitions. (Only a few relative measures using APE (Absolute Percentage Errors) are capable of both reporting results and making comparisons). Measures allowing comparisons among methods are mainly employed by academicians while those aimed at reporting or using results are used by both practitioners and academics. In Table 5 we classify the various measures in terms of their intended use and provide an indication of the extent to which we believe these measures are being used by practitioners and academics (five stars implies heaviest use, one star implies least use).

Table 5
Accuracy Measures: Type and Extent of Use
(*** = Heaviest Use ... * = Least Use)**

To Report or Use the Results of Forecasting Methods		To Make Comparisons (Evaluations) Between and Among Methods	
MSE	*****	RANKS	****
MAPE	*****	% Better	****
MAE	***	dMAPE	***
MdAPE	**	Theil's-U	**
GMMSE	*	Batting Average	**
		GMRAE	*
		MdRAE	*
		MAPE	****
		MdAPE	*

Figure 25 shows the statistical properties of reliability and discrimination for the six measures used in reporting or using the results of forecasting methods. It indicates that practically all measures (with the exception of GMMSE) are on the efficiency frontier - making the selection of a single one impossible from a statistical point of view. Figure 26 shows the "Within" and "Between" variation for the seven statistical measures aimed at comparisons among methods. It suggests that Batting Average, RANKS and dMAPE are on the statistical frontier with Theil's-U not being far from it. The remaining three measures (MdRAE, % Better and GMRAE) are on another level of their own -- exhibiting more "Within" but less "Between" than some of the methods on the efficiency frontier. Studying Figures 25 and 26 in conjunction to Table 4 can help us better understand why some measures are used more than others in practice and allow us to decide whether or not some additional methods deserve to be used to a greater extent (or why they are not used as much).

2.2.2. *Intuitiveness*

Some accuracy measures are easier to understand than others. The MAPE is intuitive as "percentage errors" are part of the everyday language. This means that it is not difficult to explain that MAPE requires "summing all absolute percentage errors" and then computing their average. This is more so as percentage errors are often reported in mass media and form an integral part of accounting and investment reports.

FIGURE 25

WITHIN AND BETWEEN VARIATION: ALL METHODS/HORIZONS MEASURES FOR A SINGLE METHOD (Six Outliers Excluded)

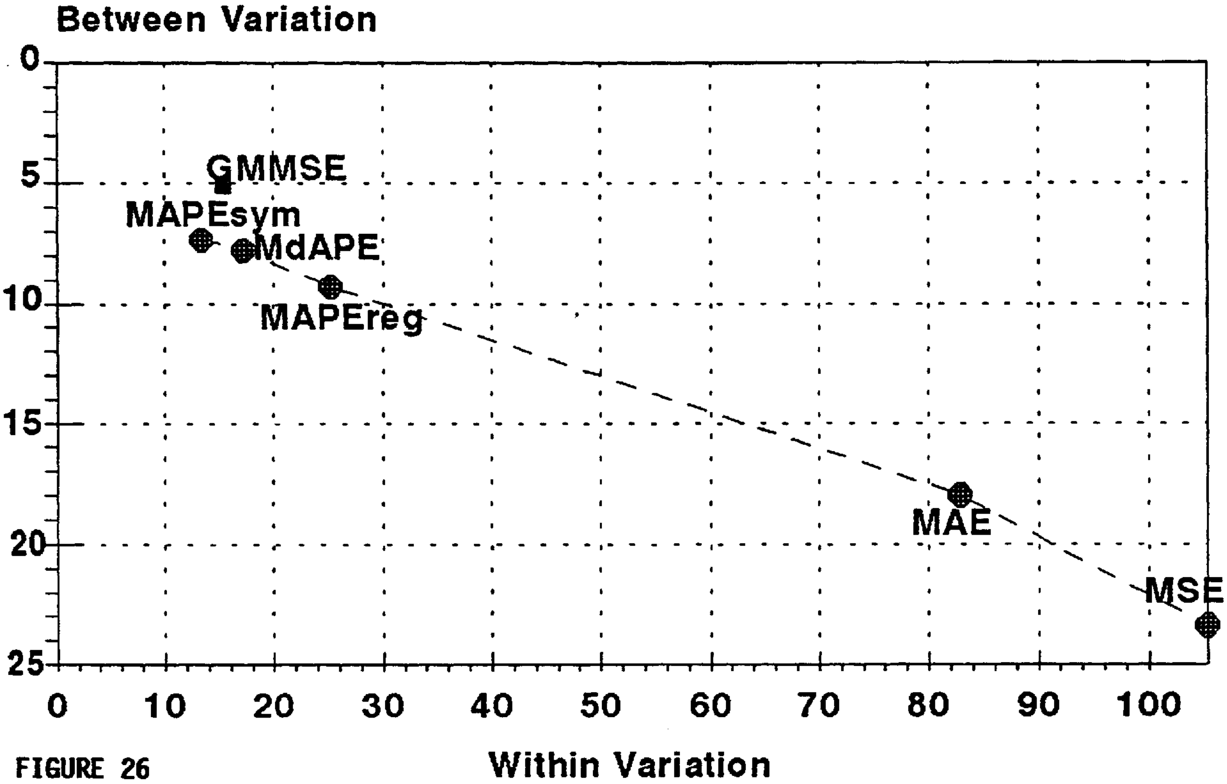
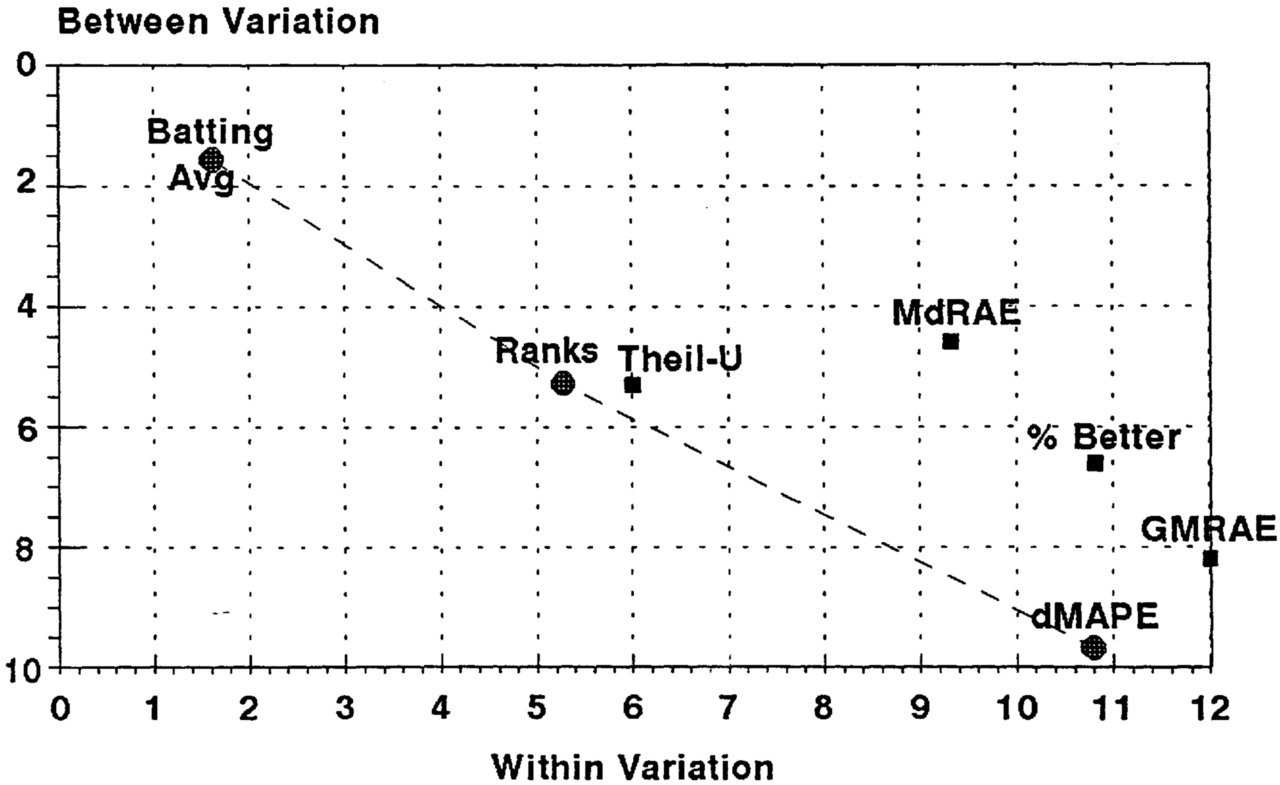


FIGURE 26

WITHIN AND BETWEEN VARIATION: ALL METHODS/HORIZONS MEASURES COMPARING TO A BENCHMARK OR OTHER METHOD(S) (Six Outliers Excluded)



Thus a percentage error of 5%, a return in investment of 8%, or that Fund XYZ outperformed the S&P500 by 8.2% over the last three years can be easily understood by the majority of people. On the other extreme stands the geometric mean square error which has little or no intuitive meaning (even to people familiar with statistics) as it involves both square terms, products and high power roots. Moreover, the GMMSE (or the MSE in this sense) are absolute measures that do not allow comparisons among methods, series and/or forecasting horizons as they heavily depend upon the absolute metric being used (that is, expressing the units in millions vs thousands would greatly affect the values of MSE and GMMSE). What does a GMMSE of 25.8 imply and how much better or worse is this compared to another one of 329.2 that refers to another series? No answer is possible. Obviously some measures like the MSE, and the MAE to a lesser extent, can be used extensively although they are not as intuitive as MAPE. However, using MSE or MAE requires that there are no other, more intuitive alternatives.

Table 6 classifies in three categories ("Common Sense Meaning", "Some Intuitive Meaning", and "Little or No Intuitive Meaning") our own appreciation of the intuitiveness of the thirteen accuracy measures used in this study. Obviously, measures with common sense meaning are, and are being preferred to those with little or no intuitive meaning.

Table 6

Classifying the Various Accuracy Measures According to their Intuitiveness

Common Sense Meaning	Some Intuitive Meaning	Little or No Intuitive Meaning
MAPE % Better dMAPE	RANKS Batting Average MdAPE MAE	MSE Theil's-U GMRAE MdRAE GMMSE

Finally, Table 7 combines the two statistical and two user oriented criteria. Ideal measures are those of high reliability and high discriminatory power which can be used for both reporting or using results and which have, in addition, common sense meaning. As there are no measures which fulfill all those criteria trade-offs must be made. In these trade-offs APE based measures figure well as they are both intuitive and can be used for reporting results and making comparisons. MSE, on the other hand, are extensively

Table 7
Classifying the Various Accuracy Measures According to the Statistical and User Oriented Criteria
(*** = Highest Usage ... * = Least Usage)**

			Statistical Criteria					
			Reliability			Discrimination		
			High	Medium	Low	High	Medium	Low
User Oriented Criteria	Informative-ness (and usage)	Reporting or Using Results		$MAPE_{sym}$ *****	MSE ***** $MAPE_{reg}$ ***** MAE *** MdAPE ** GMMSE *	MSE ***** $MAPE_{reg}$ ***** MAE *** MdAPE **	$MAPE_{sym}$ *****	GMMSE *
		Making Comparisons	Ranks **** Theil's-U ** Batting Avg **	% Better *** dMAPE ** GMRAE * MdRAE * $MAPE_{sym}$ ***	$MAPE_{reg}$ *** MdAPE *	dMAPE *** GMRAE * $MAPE_{reg}$ *** MdAPE *	% Better **** Theil's-U ** $MAPE_{sym}$ ***	Ranks **** % Better **** GMRAE *
	Intuitiveness (and understanding ability)	Common Sense Meaning		$MAPE_{sym}$ ***** % Better **** dMAPE ***	$MAPE_{reg}$ *****	$MAPE_{reg}$ ***** dMAPE ***	$MAPE_{sym}$ ***** % Better ****	
		Some Intuitive Meaning	Ranks **** Batting Avg **		MAE *** MdAPE **	MAE *** MdAPE **		Ranks **** Batting Avg ** MdRAE *
		Little or No Intuitive Meaning	Theil's-U **	GMRAE * MdRAE *	MSE ***** GMMSE *	MSE ***** GMMSE *	Theil's-U **	GMMSE *

used because of their power to discriminate among models/methods and their uniqueness of employing a quadratic loss function which is the only way of measuring the uncertainty surrounding our forecasts -- an indispensable input for inventory, material planning and other models (e.g., cash flow and financial ones).

3. DISCUSSION AND DIRECTIONS FOR FUTURE RESEARCH

This paper has shown that there is no such thing as a best statistical accuracy measure. The various measures reported in the forecasting literature are unique while, at the same time, involving trade-offs as far as the statistical criteria of reliability and discrimination as well as the user oriented ones of informativeness and intuitiveness are concerned. This means that a consensus measure that correlates as much as possible with other measures (see Armstrong and Collopy, 1992) would imply redundancy as some method(s) would be providing the same information and they would not be needed. Instead useful methods should deliver as diverse and independent information as possible. This is the biggest advantage of MSE, for instance, which correlates the least with all other methods while Theil's-U Statistic, the Batting Average, the GMRAE and MdRAE although highly related provide similar information and a false sense of consensus when their forecasts are correlated with those of other less similar methods. Thus, we believe that high correlations between or among statistical measures is a disadvantage rather than an advantage while there are no "perfect" measures, obliging us, by necessity, to make tradeoffs.

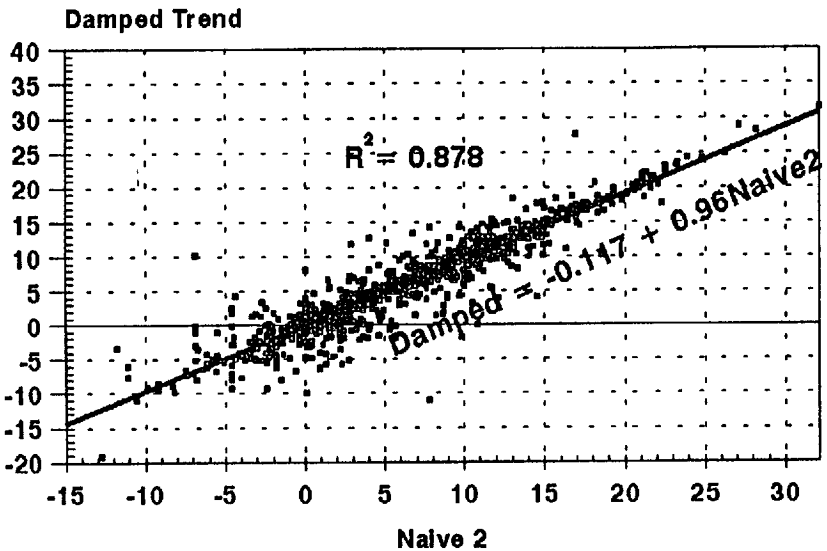
The measures of Theil's-U, Batting Average, and GMRAE should not be compared to Random Walk (Naive 1) but rather to Deseasonalized Random Walk (Naive 2) when the data is seasonal. If the benchmark is Naive 1 it is easy to prove the superiority of a method when actually such superiority is the simple outcome of the seasonality in the data which practically all methods can capture accurately. Thus, the real value of a given method needs to be compared to how much such a method can improve over and above that of seasonality which can be predicted correctly in a routine and easy manner. However, if a method is compared to Naive 2 their forecasts are often similar causing large values for these measures. Winsorizing these values by setting an arbitrary upper limit of 10, in say GMRAE, creates a serious problem when many such values exist (e.g., there were 22 tens for the 6th forecasting horizon with the 1001 series of the M-Competition when the benchmark was the Naive 2 and the method was Damped smoothing). Thus, the value of GMRAE was 0.84 when these 22 tens were included. However, when the arbitrary upper limit was set to 2 (a more logical value making the GMRAE fluctuate from 0 to 2, with 1 the base) there were 110 values set at the upper limit of 2 (a rather large number) reducing the value of GMRAE to 0.84 from 0.77, a reduction of 8.3%. For the same horizon for Holt's smoothing GMRAE value was reduced from 0.87 to 0.75 (a reduction of 13.8%).

Choosing accuracy measures should also depend on the person using them. Statisticians and others with a quantitative background have no problem using MSE, MAE, Medians or even Geometric means. However, measures intended for the general public should be restricted to those having common sense meaning, unless there is no other choice. For these reasons we recommend the MSE for statisticians once, however, something is done to deal with outliers. For non-statisticians we suggest the symmetric MAPE which is intuitive, is not influenced by outliers and can be used to both report results as well as making comparisons among methods. Moreover, from a statistical point of view $MAPE_{sym}$ is appropriate being in the middle of both reliability and discrimination (see Table 7). As the $MAPE_{sym}$ is not influenced by outliers we believe it should always be preferred to $MAPE_{reg}$ which can be influenced to a great extent by extreme values (see Figure 21 vs Figure 23).

An alternative to using benchmark methods such as the Theil's-U, Batting Average, or GMRAE is to compute the beta coefficients of a regression where the independent variable is the errors of Naive 2 (or any other benchmark method) and the dependent variable is the method being compared. The advantage of regression is that different loss functions can be used (quadratic, linear, APE, etc.) and their coefficient computed. In addition, regression can deal with outliers and allows us to know not only an overall measure such as that of Theil's-U, Batting Average, or GMRAE but also the comparison of the method in relation to specific values of the benchmark. Such a comparison, for instance, may indicate that the benchmark should be preferred for low errors while the methods for larger ones. Figure 27 shows the regression for the APE for horizon 1, 6 while Figure 28 displays the regression for the log of square errors for the same horizons when the dependent variable is Dampen smoothing and the independent Naive 2. The regression results suggest using Naive 2 when its errors are small and Dampen for larger ones. Similar regressions can help us better evaluate the performance of various methods versus some benchmarks and help us decide under what conditions it is profitable to utilize a forecasting method.

FIGURE 27(a)

NAIVE 2 vs DAMPED TREND: 1st F/C HORIZON
LOGS OF SQUARE ERRORS



NAIVE 2 vs DAMPED TREND: 6th F/C HORIZON
LOGS OF SQUARE ERRORS

FIGURE 27(b)

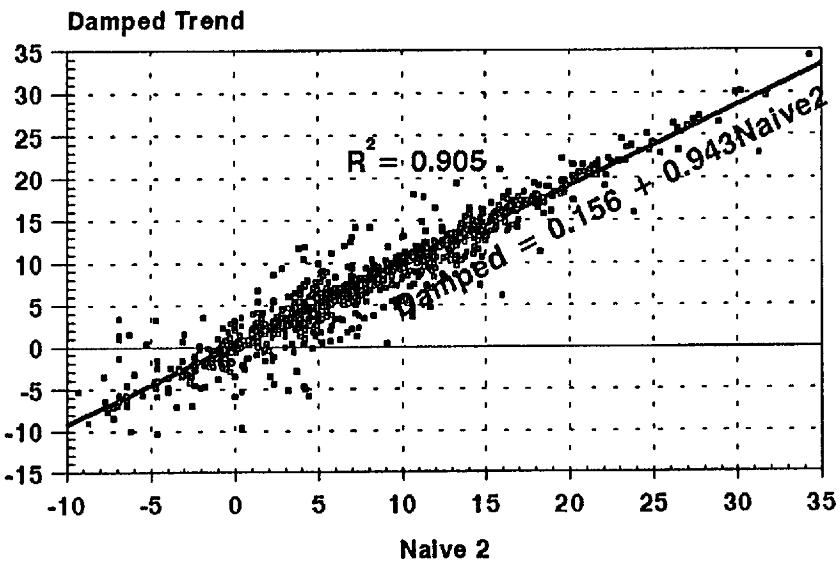
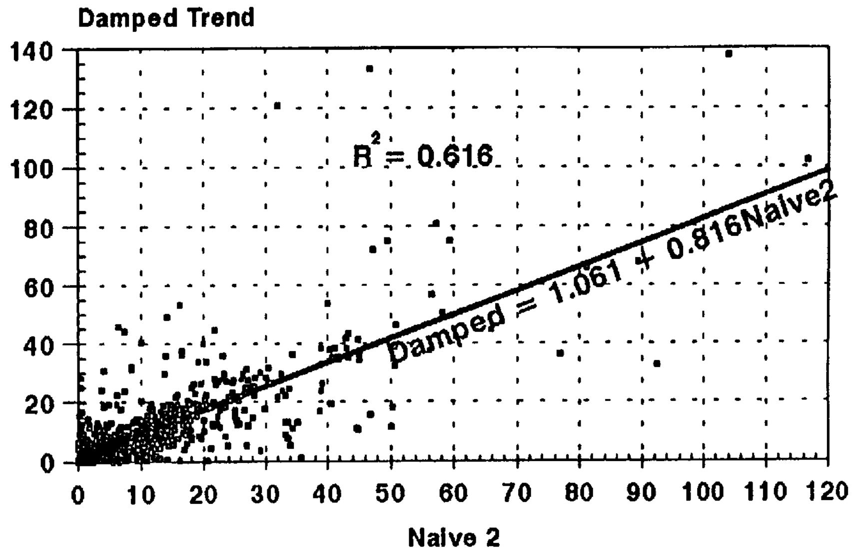


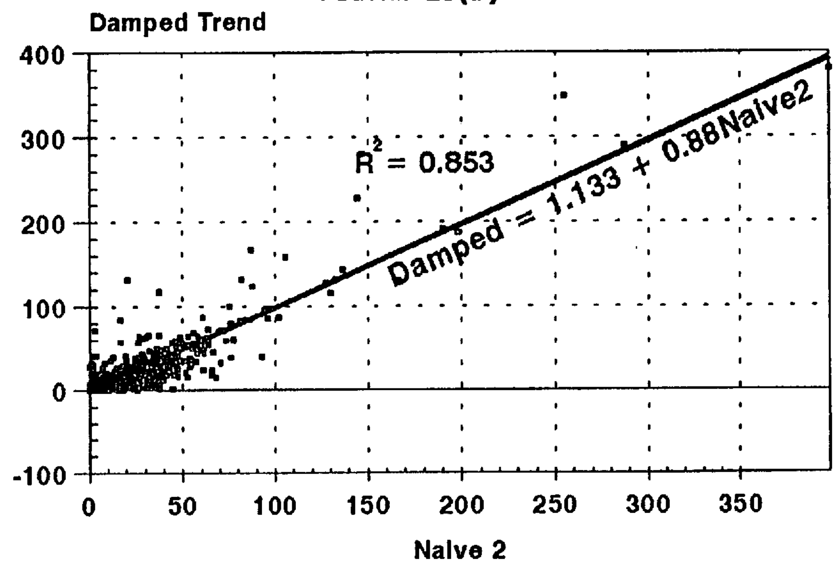
FIGURE 28(a)

NAIVE 2 vs DAMPED TREND: 1st F/C HORIZON
ABSOLUTE PERCENTAGE ERRORS



NAIVE 2 vs DAMPED TREND: 6st F/C HORIZON
ABSOLUTE PERCENTAGE ERRORS (1 Outlier Excluded)

FIGURE 28(b)



Moreover, it helps us identify and deal with outliers whose influence we can see when making a scatter diagram of the data and by computing the beta coefficients with and without the outliers.

Another direction for future research is the possibility of devising theoretical probability distribution (similar to the standard error) that will describe the post-sample "Within" and "Between" fluctuation of the various measures as a function of the sample size.

In this study we have found out that the "Within" and "Between" variations behave in a smooth fashion depending upon the standard deviation of the series, the number of series involved in each subsample, as well as the forecasting horizon and the specific method involved. It may be, therefore, possible to derive the equivalent of the standard error for various measures so that their theoretical "Within" and "Between" variations can be computed for post-sample situations. The benefits will be huge if such a task can be accomplished, providing us with invaluable help in deciding which accuracy measure to use in various situations (number of series, or sample size, forecasting horizons etc.) and forecasting methods.

Conclusions

This paper has surveyed all major accuracy (or error) measures found in the forecasting literature and evaluated them using two statistical and two user oriented criteria. The paper has shown that each of these measures provides some useful information that makes it unique. Thus, selecting among them depends upon the situation involved and the needs of decision or policy makers. Most importantly such a choice cannot be made without trade-offs. We have concluded the paper by suggesting the symmetric MAPE as a measure for both reporting results and making comparisons among methods, the dMAPE for comparing a method to a benchmark and the MSE (once possible outliers have been dealt with) for selecting appropriate models for single series and using its value to quantify the uncertainty in our predictions.

REFERENCES

Brown, R.G., (1962) *Smoothing, Forecasting and Prediction*, Prentice-Hall, Englewood Cliffs, N.J.

Carbone, R. and Armstrong, J.S., (1982) "Note: Evaluating of Extrapolative Forecasting Methods: Results of a Survey of Academicians and Practitioners", *Journal of Forecasting*, **1**, 2, 215-217.

Chatfield, C., (1988) "What is the 'best' method of forecasting?" *Journal of Applied Statistics*, **15**, 19-38.

Collopy, F. and Armstrong, J.S., (1992) "Rule-based forecasting", *Management Science*, **38**, 1394-1414.

Fildes R., (1992) "The evaluation of extrapolative forecasting methods (with discussion)",*International Journal of Forecasting*, **8**, 81-111.

Makridakis, S. et al., (1982) "The Accuracy of Extrapolative (Time Series Methods): Results of a Forecasting Competition", *Journal of Forecasting*, Vol. 1, No. 2, pp. 111-153 (lead article).

Makridakis, S., (1986) "The art and science of forecasting: an assessment and future directions", *International Journal of Forecasting*, **2**, 15-39.

McLaughlin, R.L., (1975) "The Real Record of Economic Forecasters", *Business Economics*, **10**, 3, 28-36.

Winkler, R.L., and Murphy, A.H., (1992) "On Seeking a Best Performance Measure or a Best Forecasting Method", *International Journal of Forecasting*, **8**, 1, 104-107.

Pant, P.N., and Starbuck, W.H., (1990) "Innocents in the Forecast: Forecasting and Research Methods", *Journal of Management*, **16**, 433-460.

Theil, H., (1966) *Applied Economic Forecasting*, North-Holland Publishing Co., Amsterdam.

Zellner, A., (1986) "A tale of forecasting 1001 series: The Bayesian knight strikes again",*International Journal of Forecasting*, **2**, 491-494.