

# Assessing Forecast Accuracy Measures

Zhuo Chen  
Department of Economics  
Heady Hall 260  
Iowa State University  
Ames, Iowa, 50011  
Phone: 515-294-5607  
Email: zchen@iastate.edu

Yuhong Yang  
Department of Statistics  
Snedecor Hall  
Iowa State University  
Ames, IA 50011-1210  
Phone: 515-294-2089  
Fax: 515-294-4040  
Email: yyang@iastate.edu

March 14, 2004

## Abstract

This paper looks into the issue of evaluating forecast accuracy measures. In the theoretical direction, for comparing two forecasters, only when the errors are stochastically ordered, the ranking of the forecasts is basically independent of the form of the chosen measure. We propose well-motivated Kullback-Leibler Divergence based accuracy measures. In the empirical direction, we study the performance of several familiar accuracy measures and some new ones in two important aspects: in terms of selecting the known-to-be-better forecaster and the robustness when subject to random disturbance. In addition, our study suggests that, for cross-series comparison of forecasts, individually tailored measures may improve the performance of differentiating between good and poor forecasters.

*Keywords:* Accuracy Measure, forecasting competition

### *Biographies:*

*Zhuo Chen* is Ph.D. candidate in the Department of Economics at Iowa State University. He received his BS and MS degrees in Management Science from the University of Science and Technology of China in 1996 and 1999 respectively. He graduated from the Department of Statistics at Iowa State University with MS degree in May, 2002.

*Yuhong Yang* (Corresponding author) received his Ph.D. in Statistics from Yale University in 1996. Then he joined the Department of Statistics at Iowa State University as assistant professor and became associate professor in 2001. His research interests include nonparametric curve estimation, pattern recognition, and combining procedures. He has published papers in statistics and related journals including *Annals of Statistics*, *Journal of the American Statistical Association*, *Bernoulli*, *Statistica Sinica*, *Journal of Multivariate Analysis*, *IEEE Transaction on Information Theory*, *International Journal of Forecasting* and *Econometric Theory*.

# 1 Introduction

Needless to say, forecasting is an important task in modern life. With many different methods in forecasting, understanding their relative performance is critical for more accurate prediction of the quantities of interest. Various accuracy measures have been used in the literature and their properties have been discussed to some extent. A fundamental question is: are some accuracy measures better than others? If so, in which sense? Addressing such questions is not only intellectually interesting, but also highly relevant to the application of forecasting. Not surprisingly, it is a commonly accepted wisdom that there cannot be any single best forecasting method or any single best accuracy measure, and that assessing the forecasts and the accuracy measures is necessarily subjective. However, can there be certain degree of objectivity? Obviously, it is one thing that no accuracy measure dominates the others and it is another that all reasonable accuracy measures are equally fine.

A difficulty in assessing forecast accuracy is that when different forecasts and different forecast accuracy measures are involved, the comparison of forecasts and the comparison of accuracy measures are very much entangled. Is it possible to separate these two issues?

In this work, having the above questions in mind, we intend to go one-step further both theoretically and empirically on assessing forecast accuracy measures. In the theoretical direction, when two forecasts have error distributions stochastically ordered, then the two forecasts can be compared basically regardless of the choice of the accuracy measure; on the other hand, when the forecast errors are not stochastically ordered (as is much more often the case in application), which forecast is better depends on the choice of the accuracy measure and then in general the comparison of different forecasts cannot be totally objective. As will be seen, the first part of this fact can be used to objectively compare different accuracy measures from a certain appropriate angle. If one has a good understanding of the distribution of the future uncertainty, we advocate the use of the Kullback-Leibler divergence based measures. For cross-series comparison, we argue that there can be advantage using different accuracy measures for different series. We demonstrate this advantage with several examples. In the empirical direction, we compare the popular accuracy measures and some new ones in terms of their ability to select the better forecast as well as in terms of the stability of the measures with slight perturbation of the original series. As will be seen, such forecast comparisons provide us very useful information about the behaviors of the different measures.

In the rest of the introduction we briefly review some previous related works in the literature. More details of the existing accuracy measures will be given in Sections 3 and 4.

Econometricians and Statisticians have constructed various accuracy measures to evaluate and rank forecasting methods. Diebold & Mariano (1995) proposed tests of the null hypothesis that there is no difference in accuracy between two competing forecasts. Christoffersen & Diebold (1998) suggested a forecast accuracy measure that can value the maintenance of cointegration relationships among variables. It is generally agreed that the mean squared error (Henceforth MSE) or MSE based accuracy measures

are not good choices for cross-series comparison since they are typically not invariant to scale changes. Armstrong & Fildes (1995) suggested no single accuracy measure would be the best in the sense of capturing necessary complexity of real data. This, of course, does not mean that one can arbitrarily choose a performance measure that meets a basic requirement (e.g., scale invariance). It is desirable to compare different accuracy measures to find out which measures perform better in what situations and which ones have very serious flaws and thus should be avoided in practice. We notice that only a handful of studies compared multiple forecast accuracy measures (e.g., Tashman, 2000; Makridakis 1993, Yokum and Armstrong 1995). Tashman (2000) and Koehler (2001) discussed the results of the latest M-Competition (Makridakis & Hibon, 2000) focusing on forecast accuracy measures.

The comparison of different performance measures is a very challenging task since there is no obvious way to do it objectively. To our knowledge, there has not been any systematically empirical investigation in this direction in the literature. In this work, we approach the problem from two angles: the ability of a measure to distinguish between good and bad forecasts and the stability of the measure when there is a small perturbation of the data.

Section 2 of this paper studies the theoretical comparability of different forecasts for one series and provides the theoretical motivation for the new accuracy measures. Section 3 reviews accuracy measures for cross-series comparison and we show an advantage of the use of individually tailored accuracy measures. In Section 4 we give details of the accuracy measures investigated in our empirical study. The comparison results are given in Section 5. Conclusions are in Section 6.

## 2 Theoretical comparability of different forecasts for a single series

Suppose that we have a time series  $\{Y_i\}$  to be forecasted and there are two forecasters (or two methods) with forecasts  $\hat{Y}_{i,1}$  and  $\hat{Y}_{i,2}$  of  $Y_i$  made at time  $i - 1$  based on the series itself up to  $Y_{i-1}$  and possibly with outside information available to the forecasters (such as exogenous variables). The forecast errors are  $e_{i,1} = \hat{Y}_{i,1} - Y_i$  and  $e_{i,2} = \hat{Y}_{i,2} - Y_i$  for the two forecasters respectively.

A fundamental question is how should the two forecasters be compared? Can we have any objective statement on which forecaster is doing a better job?

There are two types of comparisons of different forecasts. One is theoretical and the other is empirical. For a theoretical comparison, assumptions on the nature of the data (i.e., data generating process) must be made. But such assumptions are not needed for empirical comparisons, which draw conclusions based on data.

In this section, we consider the issue of whether two forecasters can be compared fairly. We realize the complexity of this issue and will focus our attention on a very simple setting where some theoretical understanding is possible. Basically, under a simplifying assumption on the forecast errors, we show that sometimes the two forecasts can be ordered consistently in terms of prediction risk under any reasonable loss function; for other cases, the conclusion regarding which forecaster is better depends subjectively

on the loss function chosen (i.e., it can happen that forecaster one is better under one loss function but forecaster two is better under another loss function). For the latter case, clearly, unless one can justify a particular loss function (or certain type of losses) as the only appropriate one for the problem, there is no completely objective ordering of the two forecasters.

Let the cumulative distribution functions of  $|\hat{Y}_{i,1} - Y_i|$  and  $|\hat{Y}_{i,2} - Y_i|$  be  $F_1$  and  $F_2$  respectively. Obviously, the supports of  $F_1$  and  $F_2$  are contained in  $[0, \infty)$ .

Following the statistical decision theory framework, we usually use a loss function for comparing estimators or predictions. Let  $L(Y, \hat{Y})$  be a chosen loss function. Here we only consider loss functions of the type  $L(Y, \hat{Y}) = g(|Y - \hat{Y}|)$  for a nonnegative function  $g$  defined on  $[0, \infty)$ . This class contains the familiar losses such as absolute error loss and squared error loss.

Given a loss function  $g(|Y - \hat{Y}|)$ , we say that forecaster 1 is (theoretically) better (equal or worse) than forecaster 2 if  $Eg(|e_{i1}|) < Eg(|e_{i2}|)$  ( $Eg(|e_{i1}|) = Eg(|e_{i2}|)$  or  $Eg(|e_{i1}|) > Eg(|e_{i2}|)$ ), where the expectation is with respect to the true data generating process (assumed for the theoretical investigation). Note that, given a loss function, two forecasts  $\hat{Y}_{i,1}$  and  $\hat{Y}_{i,2}$  can always be compared by the above definition at each time  $i$ .

Clearly, when multiple periods are involved, to compare two forecasters in an overall sense, assumptions on the errors are necessary. One simple assumption is that for each forecaster, the errors at different times are independent and identical distributed. Then the theoretical comparison of the forecasters is simplified to the comparison at any given time  $i$ .

In reality, however, the forecast errors are typically not iid and the comparison between the forecasters becomes theoretically intractable. Indeed, it is quite possible that forecaster 1 is better than forecaster 2 for some sample sizes but worse for other sample sizes. Even though the results in this section do not address such cases, we hope that the insight gained under the simple assumption can be helpful more generally.

## 2.1 When the forecasting error distributions are stochastically ordered

Can two forecasters be theoretically compared independently of the loss function chosen? We give a result more or less in that direction.

Here we assume that  $F_1$  is stochastically smaller than  $F_2$ , i.e., for any  $x \geq 0$ ,  $F_1(x) \geq F_2(x)$ . This means that the absolute errors of the forecasters are ordered in a probabilistic sense. It is then not surprising that the loss function does not play any important role in the theoretical comparison of the two forecasters.

**Definition:** A loss function  $L(Y, \hat{Y}) = g(|Y - \hat{Y}|)$  is said to be monotone if  $g$  is a non-decreasing function.

**Proposition 1:** If the error distributions satisfy that  $F_1$  is stochastically smaller than  $F_2$ , then for any monotone loss function  $L(Y, \hat{Y}) = g(|Y - \hat{Y}|)$ , forecaster 1 is (theoretically) no worse than forecaster 2.

The proof of Proposition 1 is not difficult and thus omitted.

From the proposition, when the error distributions are stochastically ordered, regardless of the loss function (as long as being monotone), the forecasters are consistently ordered. Therefore there is an objective ordering of the two forecasters.

Let us comment briefly on the stochastic ordering assumption. For example, if the forecast errors of forecaster 1 and 2 are both normally distributed with mean zero but different variances. Then the assumption is met. More generally, if the distributions of  $|\hat{Y}_{i,1} - Y_i|$  and  $|\hat{Y}_{i,2} - Y_i|$  both fall in a scale family, then they are stochastically ordered, and thus the forecasters are comparable naturally without the need of specifying a loss function.

However, the situation is quite different when the forecasting error distributions are not stochastically ordered, as we will see next.

## 2.2 When the forecasting error distributions are not stochastically ordered

Suppose that  $F_1$  and  $F_2$  are not stochastically ordered, i.e., there exists  $0 < x_1 < x_2$  such that  $F_1(x_1) > F_2(x_1)$  and  $F_1(x_2) < F_2(x_2)$ .

**Proposition 2:** When  $F_1$  and  $F_2$  are not stochastically ordered, we can find two monotone loss functions  $L_1(Y, \hat{Y}) = g_1(|Y - \hat{Y}|)$  and  $L_2(Y, \hat{Y}) = g_2(|Y - \hat{Y}|)$  such that forecaster 1 is better than forecaster 2 under loss function  $g_1$  and forecaster 1 is worse than forecaster 2 under loss function  $g_2$ .

Thus, from the Proposition, in general, there is no hope to order the forecasters objectively. The relative performance of the forecasts depends heavily on the loss function chosen. The proof of Proposition 2 is left to the reader.

## 2.3 Comparing forecast accuracy measures based on stochastically ordered errors

An important implication of Proposition 1 is that it can be used to objectively compare two accuracy measures from an appropriate angle. The idea is that when the errors from two forecasts are stochastically ordered, then one forecast is better than another, independently of the loss function. Consequently, we can compare the accuracy measures through their ability to pick the better forecast. This is a basis for the empirical comparison in Section 5.1.

## 2.4 How should the loss function be chosen for comparing forecasts for one series?

From the section 2.2, we know that generally, in theory, we cannot avoid the use of a loss function to compare forecasts. In this subsection, we briefly discuss the issue of choosing a loss function for comparing forecasts for one series. The issue of cross-series comparison will be addressed in Section 3.

There are different approaches. One is to use a familiar and/or mathematically convenient loss function such as squared error loss and absolute error loss. Squared error loss seems to be the most popular in statistics for mathematical convenience. Another approach is to use an intrinsic measure

which does not depend on transformations of the data. For this approach, one must make assumptions on the data generating process so that transformation-invariant measures can be derived, as will be seen soon. The third approach is to choose a loss function that seems most natural for the problem at hand based on non-statistical considerations (e.g., how the accuracy of the forecast may be related to the ultimate good of interest). Perhaps except few cases, there may be different views regarding the most natural loss functions for a particular problem.

## 2.5 Some intrinsic measures

Here we derive some intrinsic and new measures to compare different forecasts. They are obtained under strong assumptions on the data generating process. In a certain sense, these measures can pay a heavy price when the assumed data generating process does not describe the data well but they do have the advantage of a substantial gain of differentiating different forecasts when the assumed data generating process reasonably capture the nature of the data. In addition, even if the assumption on the data generating process is wrong, these measures are still sensible and better than MSE and absolute error because they are invariant under location-scale transformations.

### 2.5.1 The K-L based measure is optimal in certain sense

We assume that conditional on the previous observations of  $Y$  prior to time  $i$  and the outside information available,  $Y_i$  has conditional probability density of the form  $\frac{1}{\sigma_i} f(\frac{y-m_i}{\sigma_i})$ , where  $f$  is a probability density function (pdf) with mean zero and variance 1. Let  $\hat{Y}_i$  be a forecast of  $Y_i$ . We will consider an intrinsic distance to measure performance of  $\hat{Y}_i$ .

Kullback-Leibler divergence (information, distance) is a fundamentally important quantity in statistics and information theory. Let  $p$  and  $q$  be two probability densities with respect to a dominating measure  $\mu$ . Then the K-L divergence between  $p$  and  $q$  is defined as  $D(p \parallel q) = \int p \log \frac{p}{q} d\mu$ . Let  $X$  be a random variable with pdf  $p$  with respect to  $\mu$ . Then  $D(p \parallel q) = E \log \frac{p(X)}{q(X)}$ . It is well-known that  $D(p \parallel q) \geq 0$  (though it does not satisfy the triangle inequality and is asymmetric). Let  $X' = h(X)$ , where  $h$  is a one-to-one transformation. Let  $p'$  denote the pdf of  $X'$  and let  $q'$  denote the pdf of  $h(\tilde{X})$  where  $\tilde{X}$  has pdf  $q$ . An important property of the K-L divergence is its invariance under a one-to-one transformation. That is,  $D(p \parallel q) = D(p' \parallel q')$ . K-L divergence plays crucial roles in statistics, for examples, in hypothesis testing (Cover and Thomas 1991), minimax function estimation in deriving upper and lower bounds (e.g., Yang and Barron (1999) and earlier references thereof), and adaptive estimation (Barron (1987) and Yang (2000)).

We first assume that  $\sigma_i$  is known. Then with the forecast  $\hat{Y}_i$  replacing  $m_i$ , we have an estimated conditional pdf of  $Y_i$ :  $\frac{1}{\sigma_i} f(\frac{y-\hat{Y}_i}{\sigma_i})$ . The K-L divergence between  $p(y) = \frac{1}{\sigma_i} f(\frac{y-m_i}{\sigma_i})$  and  $q(y) = \frac{1}{\sigma_i} f(\frac{y-\hat{Y}_i}{\sigma_i})$  is  $D(p \parallel q) = \int f(x) \log \frac{f(x)}{f(x-\frac{(m_i-\hat{Y}_i)}{\sigma_i})} d\mu$ . Let  $J(a) = \int f(x) \log \frac{f(x)}{f(x-a)} d\mu$ . Note that the function  $J$  is well defined once we specify the pdf  $f$ . Now,  $D(p \parallel q) = J(\frac{m_i-\hat{Y}_i}{\sigma_i})$ . From the invariance property of the K-L divergence, for linear transformations of the series, as long as the forecasting methods are equi-variant

under linear transformations, the K-L divergence stays unchanged.

From the above points, it makes sense that if computable,  $J\left(\frac{m_i - \hat{Y}_i}{\sigma_i}\right)$  would be a good measure of performance. Due to that  $m_i$  and  $\sigma$  are unknown, one can naturally replace  $\sigma_i$  by an estimate  $\hat{\sigma}_i$  based on earlier data and replace  $m_i$  by the observed value  $Y_i$ . This motivates the loss function

$$L(Y_i, \hat{Y}_i) = J\left(\frac{Y_i - \hat{Y}_i}{\hat{\sigma}_i}\right).$$

Note that if  $\hat{\sigma}_i$  is location-scale invariant and the forecasting method is location-scale invariant, then  $L(Y_i, \hat{Y}_i)$  is location-scale invariant.

Now let's consider two special cases. One is Gaussian and the other is double exponential (Laplace).

For normal,  $f(y) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2)$  and consequently,  $J(a) = \frac{a^2}{2\sigma^2}$ . Then

$$L(Y_i, \hat{Y}_i) = J\left(\frac{Y_i - \hat{Y}_i}{\hat{\sigma}_i}\right) = \frac{(Y_i - \hat{Y}_i)^2}{2\hat{\sigma}_i^2}.$$

It perhaps is worth pointing out that  $E_i(Y_i - \hat{Y}_i)^2 = \sigma_i^2 + (m_i - \hat{Y}_i)^2$ , where  $E_i$  denotes expectation conditional on the information prior to observing  $Y_i$ . Since  $\sigma_i^2$  does not depend on any forecasting method, for this case,  $J\left(\frac{Y_i - \hat{Y}_i}{\hat{\sigma}_i}\right)$  is essentially equivalent to  $J\left(\frac{m_i - \hat{Y}_i}{\sigma_i}\right)$ .

For double exponential,  $f(y) = \frac{1}{2} \exp(-|y|)$  and  $J(a) = \exp(-\frac{|a|}{\sigma}) - 1 - \frac{|a|}{\sigma}$ . Then

$$L(Y_i, \hat{Y}_i) = J\left(\frac{Y_i - \hat{Y}_i}{\hat{\sigma}_i}\right) = \exp\left(-\frac{|Y_i - \hat{Y}_i|}{\hat{\sigma}_i}\right) - 1 - \frac{|Y_i - \hat{Y}_i|}{\hat{\sigma}_i}.$$

For our approach, the main difficulty is the estimation of  $\sigma_i$ , especially when dealing with nonstationary series.

## 2.5.2 The best choice of loss depends on the nature of the data

Here we show that the MSE is the optimal choice of performance measure in a certain appropriate sense when the errors have normal distributions and absolute error (ABE) is the optimal choice when the errors have double exponential distributions.

Consider two forecasts  $\hat{Y}_{i,1}$  and  $\hat{Y}_{i,2}$  of  $Y_i$  with the forecast errors  $e_{i,1} = \hat{Y}_{i,1} - Y_i$  and  $e_{i,2} = \hat{Y}_{i,2} - Y_i$  respectively. We assume that the errors are iid with mean zero for both the forecasters.

**MSE or ABE?** First we assume that  $e_{i,1} \sim N(0, \sigma_1^2)$  and  $e_{i,2} \sim N(0, \sigma_2^2)$ . From Proposition 1, we know that forecaster 1 is theoretically better than forecaster 2 if  $\sigma_1^2 < \sigma_2^2$  for any monotone loss function. In reality, of course, one does not know the variances and need to compare the forecasters empirically by looking at the history of their forecasting errors. In this last task, the choice of a loss function becomes important.

Under our assumptions,  $(e_{1,1}, \dots, e_{n,1})$  has joint pdf

$$\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^n e_{i,1}^2}{2\sigma_1^2}\right)$$

and  $(e_{1,2}, \dots, e_{n,2})$  has joint pdf

$$\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^n e_{i,2}^2}{2\sigma_2^2}\right).$$

Thus for each of the two forecasters,  $\sum_{i=1}^n e_{i,j}^2$  is a sufficient statistic for  $j = 1, 2$ . In contrast,  $\sum_{i=1}^n |e_{i,j}|$  is not sufficient. This suggests that for each forecaster, when the errors are normally distributed, the use of MSE ( $\sum_{i=1}^n e_{i,j}^2$ ) better captures the information in the errors than other choices including ABE ( $\sum_{i=1}^n |e_{i,j}|$ ). On the other hand, when the errors have double exponential distribution,  $\sum_{i=1}^n |e_{i,j}|$  is sufficient but  $\sum_{i=1}^n e_{i,j}^2$  is not, and thus the choice of ABS is better than MSE. Note also that when the errors of the two forecasts are all independent and normally distributed, for testing  $H_0 : \sigma_1^2 = \sigma_2^2$  versus  $H_1 : \sigma_1^2 \neq \sigma_2^2$ , there is a uniformly most powerful unbiased test based on  $\sum_{i=1}^n e_{i,1}^2 / \sum_{i=1}^n e_{i,2}^2$ , which again is in the form of MSE.

**A simulation** Here we study the two types of errors mentioned above.

Case 1 (normal).  $e_{i,1} \sim N(0, 1)$  and  $e_{i,2} \sim N(0, 1.5)$  for  $i = 1, \dots, 100$ . Replicate 1000 times and record the numbers of times  $\sum_{i=1}^n e_{i,1}^2 > \sum_{i=1}^n e_{i,2}^2$  and  $\sum_{i=1}^n |e_{i,1}| > \sum_{i=1}^n |e_{i,2}|$ , respectively.

Case 2 (double exponential).  $e_{i,1} \sim DE(0, 1)$  and  $e_{i,2} \sim DE(0, 1.5)$  for  $i = 1, \dots, 100$ . Replicate 1000 times and record the number of times  $\sum_{i=1}^n e_{i,1}^2 > \sum_{i=1}^n e_{i,2}^2$  and  $\sum_{i=1}^n |e_{i,1}| > \sum_{i=1}^n |e_{i,2}|$ , respectively.

The numbers of times that the above inequalities hold are presented in Table 1.

|               | Squared Error | Absolute Error |
|---------------|---------------|----------------|
| <i>Normal</i> | 0.311         | 0.327          |
| <i>DE</i>     | 0.270         | 0.247          |

Table 1: Comparing MSE and ABE

From the above simulations, we clearly see that for differentiating the competing forecasters, the choice of loss function does matter. When the errors are normally distributed, MSE is better and when the errors are double exponentially distributed, ABE is better. A sensible recommendation for application is that when the errors look like normally distributed (e.g., by examining the Q-Q plot), MSE is a good choice; and when the errors seem to have a distribution with heavier tail, ABE is a better choice.

### 3 Accuracy measures for cross-series comparison

Forecast accuracy measures have been used in empirical evaluation of forecasting methods, e.g., in the M-Competitions (Makridakis, Hibon & Moser 1979; Makridakis & Hibon 2000). Measures used in M1 Competition are: MSE (Mean square error), MAPE (Mean average percentage error), and Theil's U2-statistic. More measures are used in M3 Competition, i.e., symmetric mean absolute percentage error (sMAPE), average ranking, percentage better, median symmetric absolute percentage error (msAPE), and median relative absolute error (mRAE).

Here we classify the forecast accuracy measures into two types. The first category is stand-alone measures, i.e., measures can be determined by the forecast under evaluation alone. The second type is the relative measures that compare the forecasts to a baseline/naive forecast, i.e., random walk, or a (weighted) average of available forecasts.

### 3.1 Stand-Alone Accuracy Measures

Stand-alone accuracy measures are those that can be obtained without additional reference forecasts. They are usually associated with a certain loss function though there are a few exceptions (e.g., Granger & Jeon (2003a,b) proposed a time-distance criterion for evaluating forecasting models). In our study, we include several accuracy measures that are based on quadratic and absolute loss functions.

Accuracy measures based on mean squared error criterion, especially MSE itself, have been used widely for a long time in evaluating forecasts for a single series. Indeed, Carbone and Armstrong (1982) found that Root Mean Squared Error (RMSE) had been the most preferred measure of forecast accuracy. However, for cross-series comparison, it is well known that MSE and the like are not appropriate since they are not unit-free. Newbold (1983) criticized the use of MSE in the first M-Competition (Makridakis et al., 1982). Clements & Hendry (1993) proposed the Generalized Forecast Error Second Moment (GFESM) as an improvement to the MSE. Armstrong & Fildes (1995) again suggested that the empirical evidence showed that the mean square error is inappropriate to serve as a basis for comparison.

Ahlburg (1992) found that out of seventeen population research papers he surveyed, ten used Mean Absolute Percentage Error (MAPE). However, MAPE was criticized for the problem of asymmetry and instability when the original value is small (Koehler, 2001; Goodwin & Lawton, 1999).

In addition, Makridakis (1993) pointed out that MAPE may not be appropriate in certain situations, such as budgeting, where the average percentage errors may not properly summarize accounting results and profits. MAPE as accuracy measure is affected by four problems: (1) Equal errors above the actual value result in a greater APE; (2) Large percentage errors occur when the value of the original series is small; (3) Outliers may distort the comparisons in forecasting competitions or empirical studies; (4) MAPEs cannot be compared directly with naïve models such as random work (Makridakis 1993). Makridakis (2000) proposed modified MAPE measure (Symmetric Median Absolute Percent Error) and used it in the M2 and M3 competitions. However, Koehler (2001) found sMAPE penalizes low forecasts more than high forecasts and thus favors large predictions than smaller ones.

### 3.2 Relative Measures

The idea of relative measures is to evaluate the performance of a forecast relative to that of a benchmark (sometimes just a “naive”) forecast. Measures may produce very big numbers due to outliers and/or inappropriate modeling, which in turn make the comparison of different forecasts not feasible or not reliable. A shock may make all forecasts perform very poorly, and stand-alone measures may put excessive weight on this period and choose a measure that is less effective in most other periods. Relative

measures may eliminate the bias introduced by potential trends, seasonal components and outliers, provided that the benchmark forecast handles these issues appropriately. However, we need to note that choosing the benchmark forecast is subjective and not necessarily easy. The earliest relative forecast accuracy measure seems to be Theil’s U2-statistic, of which the benchmark forecast is the value of the last observation.

Collopy and Armstrong (1992a) suggested that Theil’s U2 had not gained more popularity because it was less easy to communicate. Collopy and Armstrong (1992b, p.71) proposed a similar measure (RAE). Thompson (1990) proposed an MSE based statistic– log mean squared error ratio– as an improvement of MSE to evaluate the forecasting performances across different series.

### 3.3 The same measure across series or individually tailored measures?

As far as we know, in cross-series comparison of different forecasters, for each measure under investigation, it is applied to all the series. A disadvantage of this approach is that a fixed measure may be well suited for some series but may be inappropriate for others (e.g., due to a lack of power to distinguish different forecasts or too strong influence by a few points). For such cases, using individually tailored measures may improve the comparison of the forecasters.

**Example 1:** Suppose that the data set has 100 series. The sample size for each series is 50. The first 75 series are generated as  $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + e$ , where  $\alpha_0, \alpha_1, \alpha_2, \alpha_3$  are generated as random draws from uniform distribution  $unif(-1, 1)$ ,  $x_1, x_2, x_3$  are exogenous variables independently distributed as  $N(0, 1)$  and  $e$  is independent and normal distributed as  $N(0, 5)$ . The remaining 25 series are generated as  $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + e$  as above except that  $e$  is distributed as double exponential  $DE(0, \sqrt{10}/2)$ .

We compare the two forecasts  $y^1$  and  $y^2$ , which are generated by:

$$\begin{aligned} y_t^1 &= \alpha_0 + \alpha_1 x_{1t} + \alpha_2 x_{2t} + \alpha_3 x_{3t}, \\ y_t^2 &= \hat{\alpha}_{0t} + \hat{\alpha}_{1t} x_{1t} + \hat{\alpha}_{2t} x_{2t} + \hat{\alpha}_{3t} x_{3t}, \end{aligned}$$

where  $\hat{\alpha}_{0t}, \hat{\alpha}_{1t}, \hat{\alpha}_{2t}, \hat{\alpha}_{3t}$  are estimated adaptively by regressing  $y$  on  $x_1, x_2$  and  $x_3$  (with a constant term) on previous data, i.e.,  $x_{1,1}, x_{1,2}, \dots, x_{1,t-1}; x_{2,1}, x_{2,2}, \dots, x_{2,t-1}; x_{3,1}, x_{3,2}, \dots, x_{3,t-1}$ . Note that  $y_t^1$  is the “ideal” forecast with the parameters known.

We consider three measures to compare the two forecasts. One is the KL-N , another is the KL-DE2<sup>1</sup>, and the third is an adaptive measure that uses KL-N for the first 75 series and KL-DE2 for the remaining 25 series. The two forecasts are evaluated based on their forecasts of the last 10 periods. We make 2000 replications and record the percentage of choosing the better forecast<sup>2</sup> (i.e.,  $y_t^1$ ) by the three measures.

We report the means and their corresponding standard errors of the difference between the percentage of choosing the better forecast by the individually tailored measure and the other two measures in Table

<sup>1</sup>Please refer to the next section for the details of KL-N and KL-DE2.

<sup>2</sup>We understand that there might be concerns over whether the conditional mean is ideal or not, but it is definitely free of estimation error. Furthermore, since we are varying the coefficients for each series and average the percentage over the 100 series, we pretty much eliminate the possibility that  $y_t^2$  “happens” to be superior to the conditional mean.

2. The table shows that the individually tailored measure improves the ability to distinguish between forecasts with different accuracy. The improvement of the percentage of choosing the better forecast is about 0.19% to the KL-DE2 and 0.58% to KL-N. Besides being statistically significant, even though these numbers seem to be small, they are not practically insignificant (note that Makridakis & Hibon (2000) showed that the percentage better of sixteen forecasting procedures with respect to a baseline method was from -1.7% to 0.8%).

|                                      | <i>KL - N</i> | <i>KL - DE2</i> | Adaptive Measures |
|--------------------------------------|---------------|-----------------|-------------------|
| Example 1                            |               |                 |                   |
| Percent                              | 71.60%        | 71.99%          | 72.18%            |
| Difference with the Adaptive measure | 0.58%         | 0.19%           |                   |
| Standard error of the difference     | 0.03%         | 0.05%           |                   |
| Example 2                            |               |                 |                   |
| Percent                              | 65.00%        | 72.90%          | 73.00%            |
| Difference with the Adaptive measure | 1.30%         | 0.14%           |                   |
| Standard error of the difference     | 0.04%         | 0.04%           |                   |
| Example 3                            |               |                 |                   |
| Percent                              | 81.11%        | 81.18%          | 81.68%            |
| Difference with the Adaptive measure | 0.57%         | 0.50%           |                   |
| Standard error of the difference     | 0.04%         | 0.03%           |                   |

Table 2: Percentage of Choosing the Better Forecast

**Example 2:** Example 2 has the same setting as in Example 1 except that we change the ratio of series with normal error and double exponential error to 1:1. The new measure is still better than that of the two original measures but the extent varies, which gives another evidence that the performance of accuracy measures may be influenced by the error structure.

**Example 3:** To address the concern that the conditional mean may not necessarily be better than the other forecast, we generate  $y$  as a series random drawn from a uniform distribution is  $unif(0, 1)$  and the two forecasts are:  $y^1 = y + e_1$ ,  $y^2 = y + e_2$ , where  $e_{1t}$  is distributed as iid  $N(0, 1)$  and  $e_{2t}$  is distributed as iid  $N(0, 2)$  for the first 50 series and  $e_{1t}$  is distributed as iid  $DE(0, \sqrt{2}/2)$  and  $e_{2t}$  is distributed as iid  $DE(0, 2)$  for the remaining 50 series. Replicate it for 2000 times and we report the quantities in the lower part of Table 2. In this case, it is obvious that  $y^2$  is stochastically dominated by  $y^1$  in forecast accuracy, and thus we know for sure that  $y^1$  is the better forecast. The result is similar to those in Examples 1 and 2.

The examples show that it is potentially better to use adaptive measures (as opposed to a fixed measure) when comparing forecasts. The adaptive measure (or individually tailored measures) can better distinguish the candidate forecasts using the individual characteristics of the series. It should be mentioned that in these examples, KL-N and KL-DE2 are applied with the knowledge of the nature of the series. In a real application, of course, one is not typically told whether the forecast errors are normally distributed or double-exponentially distributed. One then needs to analyze this aspect using, e.g., Q-Q plots or formal tests. We leave this for future work.

## 4 Measures in Use in our Empirical Study

In the empirical study of this paper, we try to assess eighteen accuracy measures, including a few new ones motivated from K-L divergence.

### 4.1 Stand-Alone Accuracy measures

We consider eleven stand-alone accuracy measures. MAPE, sMAPE and RMSE are familiar in the literature. We propose several new measures based on Kullback-Leibler divergence, i.e., KL-N, KL-N1, KL-N2, KL-DE1, and KL-DE2. We also suggest several variations of MSE and APE based measures, i.e., msMAPE, NMSE. IQR is a new measure based on MSE and adjusted by inter quartile range. Let  $m$  be the number of observations we use in the evaluation of forecasts. Below we give the details of the aforementioned measures.

The commonly used MAPE (mean absolute percentage error) has the form:

$$\frac{1}{m} \sum_{i=1}^m \frac{|\hat{y}_i - y_i|}{|y_i|}.$$

Makridakis & Hibon (2000) used sMAPE (symmetric mean absolute percentage error):

$$\frac{1}{m} \sum_{i=1}^m \frac{|\hat{y}_i - y_i|}{(|y_i| + |\hat{y}_i|)/2}.$$

The measure reaches the maximum value of two when either  $|y_i|$  or  $|\hat{y}_i|$  equals to zero (undefined when both are zero).

To avoid the possibility of an inflation of sMAPE caused by zero values in the series, we add a component in the denominator of the symmetric MAPE and denote it msMAPE (modified sMAPE), which is formulated as:

$$\frac{1}{m} \sum_{i=1}^m \frac{|\hat{y}_i - y_i|}{(|y_i| + |\hat{y}_i|)/2 + S_i},$$

where  $S_i = \frac{1}{i-1} \sum_{k=1}^{i-1} |y_k - \bar{y}_{i-1}|$ ,  $\bar{y}_{i-1} = \frac{1}{i-1} \sum_{k=1}^{i-1} y_k$ .

RMSE is the usual root mean square error measure:

$$\sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2}.$$

NMSE (normalized MSE) is formulated as:

$$\sqrt{\frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2}},$$

where  $\bar{y} = \frac{1}{n} \sum_{k=1}^n y_k$ .

KL-N is proposed based on the Kullback-Leibler (KL) divergence. The measure corresponds to the quadratic loss function (normal error) scaled with (adaptively moving) variance estimate. Its formula is:

$$\sqrt{\frac{1}{m} \sum_{i=1}^m \frac{(\hat{y}_i - y_i)^2}{S_i^2}},$$

where  $S_i^2 = \frac{1}{i-1} \sum_{k=1}^{i-1} (y_k - \bar{y}_{i-1})^2$ ,  $\bar{y}_{i-1} = \frac{1}{i-1} \sum_{k=1}^{i-1} y_k$ . We discussed the theoretical motivation of K-L divergence based measures in Section 2.5.1.

KL-N1 is a modified version of KL-N. We use a different variance estimate that only considers the last 5 periods. The reason for considering only a few recent periods is to allow the variance estimator to perform well when  $S_i^2$  does not converge properly due to e.g., un-removed trends. Its formula is:

$$\sqrt{\frac{1}{m} \sum_{i=1}^m \frac{(\hat{y}_i - y_i)^2}{S_{i,5}^2}},$$

where  $S_{i,5}^2 = \frac{1}{5} \sum_{k=i-6}^{i-1} (y_k - \bar{y}_{i-1,5})^2$ ,  $\bar{y}_{i-1,5} = \frac{1}{5} \sum_{k=i-6}^{i-1} y_k$ .

KL-N2 uses a variance estimate that considers the last 10 period. Its formula is:

$$\sqrt{\frac{1}{m} \sum_{i=1}^m \frac{(\hat{y}_i - y_i)^2}{S_{i,10}^2}},$$

where  $S_{i,10}^2 = \frac{1}{10} \sum_{k=i-10}^{i-1} (y_k - \bar{y}_{i-1,10})^2$ ,  $\bar{y}_{i-1,10} = \frac{1}{10} \sum_{k=i-10}^{i-1} y_k$ .

KL-DE1 is an accuracy measure we proposed based on the K-L divergence and the assumption of double exponential error. Its formula is:

$$\frac{1}{m} \sum_{i=1}^m \left( e^{-\frac{|\hat{y}_i - y_i|}{\hat{\sigma}_i}} + \frac{|\hat{y}_i - y_i|}{\hat{\sigma}_i} - 1 \right),$$

where  $\hat{\sigma}_i^2 = \frac{1}{i-1} \sum_{j=1}^{i-1} (y_j - \bar{y}_{i-1})^2$ .

KL-DE2 is an accuracy measure similar to KL-DE1 but with a different estimator of the scale parameter from the one used in KL-DE1. Its formula is same with KL-DE1 but  $\hat{\sigma}_i = \frac{1}{i-1} \sum_{j=1}^{i-1} |y_j - \bar{y}_{i-1}|$ .

IQR is an accuracy measure based on inter quartile range. Its formula is:

$$\sqrt{\frac{1}{m} \sum_{i=1}^m \frac{(\hat{y}_i - y_i)^2}{Iqr^2}},$$

where  $Iqr$  is the inter quartile range of  $Y_1, \dots, Y_m$  defined as the difference between the third quartile and the first quartile of the data. Note that this measure is local-scale transformation invariant and normalizes the absolute error in terms of  $Iqr$ .

## 4.2 Relative Accuracy Measures

We will use seven relative forecast accuracy measures.

RSE (Relative Squared Error) is the square root of the mean of the ratio of MSE relative to that of random walk forecast at the evaluated time periods. It is motivated by RAE (relative absolute error) proposed by Collopy and Armstrong (1992b). It is formulated as:

$$\sqrt{\frac{1}{m} \sum_{i=1}^m \frac{(\hat{y}_i - y_i)^2}{(y_i - y_{i,rw})^2}},$$

where  $y_{i,rw} = y_{i-1}$ .

We propose mRSE (modified RSE) to improve RSE in the case when the series remains unchanged for one or more time periods. To achieve this, we add a variance estimates component to the denominator, thus its formula can be written as:

$$\sqrt{\frac{1}{m} \sum_{i=1}^m \frac{(\hat{y}_i - y_i)^2}{(y_i - y_{i,rw})^2 + S_i^2}},$$

where  $y_{i,rw} = y_{i-1}$ ,  $S_i^2 = \frac{1}{i-1} \sum_{k=1}^{i-1} (y_k - \bar{y}_{i-1})^2$ ,  $\bar{y}_{i-1} = \frac{1}{i-1} \sum_{k=1}^{i-1} y_k$  (an alternative is to replace  $S_{i-1}^2$  by the average of  $(y_i - y_{i,rw})^2$ ).

Theil's U2 is:

$$\sqrt{\frac{\sum (\hat{y}_i - y_i)^2}{\sum (y_i - y_{i,rw})^2}},$$

RAE (Collopy and Armstrong, 1992b) is:

$$\sqrt{\frac{\sum |\hat{y}_i - y_i|}{\sum |y_i - y_{i,rw}|}},$$

It should be pointed out that the relative measures are not without any problem. For example, if for one series, a forecasting method is much worse than random walk, then the measure can be arbitrarily large, which can be overly influential when multiple series are compared. Another weakness is that when the random walk forecast is very poor, then the measures take very small values and consequently these series play a less important role compared to series where random walk forecast is comparable to the other forecasts.

MSEr1 (MSE relative 1) is the square root of the mean of the ratio of MSE relative to the variance of the available forecasts at the current time. Its formula is:

$$\sqrt{\frac{1}{m} \sum_{i=1}^m \frac{(\hat{y}_i - y_i)^2}{\frac{1}{k} \sum_{j=1}^k (\hat{y}_{ji} - y_i)^2}},$$

where  $\hat{y}_{ji}$  is the  $j$ th forecast of  $i$ th observation.

MSEr2 (MSE relative 2) is the square root of the mean of the ratio of MSE relative to the sample variance of the difference between  $Y$  and the mean of the competing forecasts. Its formula is:

$$\sqrt{\frac{1}{m} \sum_{i=1}^m \frac{(\hat{y}_i - y_i)^2}{\frac{1}{i-1} \sum_{l=1}^{i-1} (y_l - \bar{\hat{y}}_l)^2}},$$

where  $\bar{\hat{y}}_l = \frac{1}{k} \sum_{j=1}^k \hat{y}_{jl}$ .

MSEr3 (MSE relative 3) is the square root of the mean of the ratio of MSE relative to the average mean squared errors of the candidate forecasts. Its formula is:

$$\sqrt{\frac{1}{m} \sum_{i=1}^m \frac{(\hat{y}_i - y_i)^2}{\frac{1}{k} \sum_{l=1}^k \frac{1}{i-1} \sum_{j=1}^{i-1} (y_j - \hat{y}_{lj})^2}}.$$

## 5 Evaluating the Accuracy Measures

Armstrong & Fildes (1995) pointed out that the purpose of an error measure is to provide an informative and clear summary of the error distribution. They suggested that error measure should use a well-specified loss function, be reliable, resistant to outliers, comprehensible to decision makers and should also provide a summary of the forecast error distribution for different lead times. Clements and Hendry (1993) emphasized that the robustness of an error measure to the linear transformation of the original series is an important factor to consider.

In this section we evaluate the performance of the forecast accuracy measures from two angles. We investigate the ability of the measures in picking up the “better” forecast; study the stability of the forecasts to small disturbances on the original series and the stability of the measures to linear transformations of the series.

### 5.1 Ability to select the better forecast

Naturally, we hope that a forecast accuracy measure can differentiate between good and poor forecasts. For real data sets, we cannot decide which forecast is really the best if different measures disagree and there is no dominant forecast. Part of the reason is that we have no definite information on the real data generating process (DGP).

When selecting the “better” (or “best”) forecast is the criterion, of course, defining “better” (or “best”) appropriately is crucial. However this becomes somewhat circular because an accuracy measure is typically needed to quantify the comparison between the forecasts. To overcome the difficulty, our strategies are as follows.

Suppose a forecaster is given the information of the DGP with known structural parameters. Then the conditional mean can be naturally used as a good forecast. For a forecaster who is given the form of the DGP but with unknown structural parameters, he/she needs to estimate the parameters for forecasting, which clearly introduces additional variability in the forecast. Since the first one should be advantageous compared to the second one, we can evaluate an accuracy measure in terms of its tendency to prefer the first one. Moving further in this direction, we can work with two forecasters that have stochastically ordered error distributions and assess the goodness of an accuracy measure using the frequency that it yields a better value for the better forecaster.

We agree with Armstrong & Fildes (1995) that simulated data series might not be a good representation of real data. Given a forecast accuracy measure, data sets can be used to evaluate the competing forecasts objectively. For assessing an accuracy measure, however, due to the fact that the effects of the forecasts and the accuracy measure are entangled, maintaining objective and informative is much more challenging. The use of simulated data then becomes important for a rigorous comparison of accuracy measures.

We consider nine cases in this subsection.

### 5.1.1 Cases 1-7

The seven cases in this subsection represent various scenarios we may encounter in real applications (but obviously by no means they give a complete representation) and they can give us some useful information regarding the performance of the accuracy measures. We replicate all the simulation 20000 times. The numbers reported in Table 3 are the percentages that each measure chooses the better forecast over all the replications.

1. Data generating process is AR(1) with auto-regressive coefficient 0.75, and the series length is 50. Random disturbance is distributed as  $N(0,1)$ . Using the eighteen measures, we compare the forecasts generated by the true model, in which we know the true model structure but not the structural parameters, to the better forecast available, which is the conditional mean of the series (i.e., when the auto-regression coefficient is known).

Figure 1 presents the boxplot of the values measured for the forecasts produced by the conditional mean and when  $m = 20$ . The values greater than 20 are clipped. From the figure, clearly for some of the measures, the distribution are highly asymmetric.

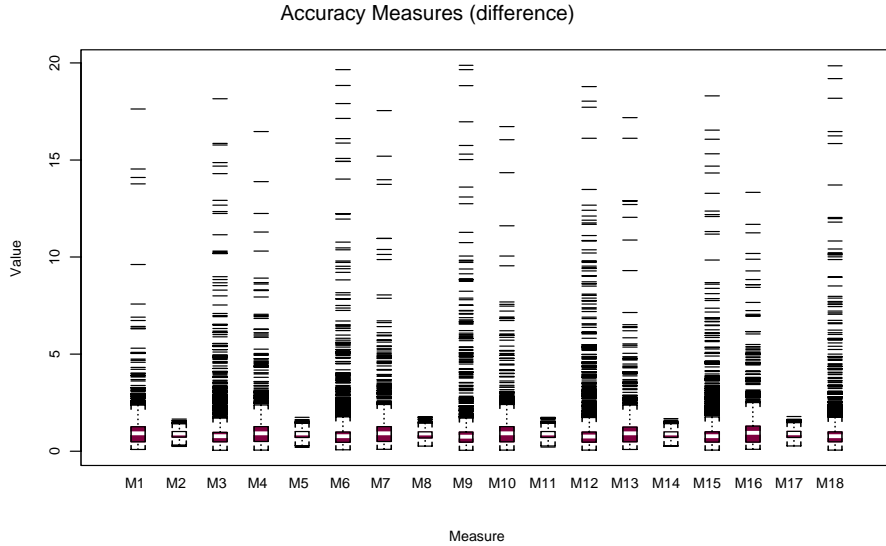


Figure 1: Boxplot of the Values of the Accuracy Measures

2. Data generating process is white noise distributed as  $N(0,1)$ , and the series length is 50. We compare forecasts generated by a white noise, which is also distributed as  $N(0,1)$ (true model) to the better forecast, which is the conditional mean of the series, zero.

3. Data generating process is AR(1) with auto-regressive coefficient 0.75, and the series length is 50. Random disturbance is distributed as  $N(0,1)$ . We compare the forecast generated by white noise process distributed as  $N(0,1)$ , to the better forecast available, i.e., the conditional mean of the series.

4. Data generating process is:

$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + e$ , where  $\alpha_0$  is generated as a random draw from uniform distribution  $unif(0, 1)$ ,  $\alpha_1, \alpha_2, \alpha_3$  are generated as random draws from uniform distribution  $unif(-1, 1)$ . The sample size  $n = 50$ ,  $x_{1t}, x_{2t}, x_{3t}$  are exogenous variables independently generated from  $N(0, 1)$  and  $e_t$  is the random disturbance distributed as iid  $N(0, 1)$ ,  $t = 1, \dots, n$ . We compare the two forecasts  $y^1$  and  $y^2$ , where  $y^1$  is generated by assuming we know the true parameters and  $y^2$  is generated using coefficients estimated based on available data:

$$\begin{aligned} y_t^1 &= \alpha_0 + \alpha_1 x_{1t} + \alpha_2 x_{2t} + \alpha_3 x_{3t} \\ y_t^2 &= \hat{\alpha}_{0t} + \hat{\alpha}_{1t} x_{1t} + \hat{\alpha}_{2t} x_{2t} + \hat{\alpha}_{3t} x_{3t} \end{aligned}$$

for  $t = n - m + 1, \dots, n$ , where  $\hat{\alpha}_{0t}, \hat{\alpha}_{1t}, \hat{\alpha}_{2t}, \hat{\alpha}_{3t}$  are estimated by regressing  $y$  on  $x_1, x_2$ , and  $x_3$  with a constant term using previous data, i.e.,  $x_{1,1}, x_{1,2}, \dots, x_{1,t-1}; x_{2,1}, x_{2,2}, \dots, x_{2,t-1}; x_{3,1}, x_{3,2}, \dots, x_{3,t-1}$ .

5. The setting of Case 5 is the same as Case 4 except that  $e_t$  is distributed as double exponential  $DE(0, 1)$  for  $t = 1, \dots, n$ .

6. Data generating process is:  $y = x_1$ , where  $x_1$  is exogenous variables independently distributed as  $unif(0, 1)$ . The sample size  $n = 50$ . We compare the two forecasts  $y^1$  and  $y^2$ , which are generated by:

$$\begin{aligned} y_t^1 &= x_{1t} + e_{1t} \\ y_t^2 &= x_{1t} + e_{2t} \end{aligned}$$

for  $t = n - m + 1, \dots, n$ , where  $e_{1t}$  distributed as iid  $N(0, 1)$  and  $e_{2t}$  distributed as iid  $N(0, 2)$ . Note that here  $y_t^1$  dominates  $y_t^2$  independently of the loss function.

7. The setting of Case 7 is the same as Case 6 except that  $e_{1t}$  distributed as iid  $DE(0, 1)$  and  $e_{2t}$  distributed as iid  $DE(0, 2)$ . As in Case 6,  $y_t^1$  beats  $y_t^2$ .

The results in Table 3 reveal the following. First, sMAPE performs very poorly when the true value is close to zero. A forecast of zero will be deemed as the worst (maximum in value) of the measured performance, no matter what values the other forecasts take. If the true value is zero, the measure will also give out the maximum error measure of 2 for any forecast not equal to zero. After adding a non-negative component to the denominator, the msMAPE is superior to sMAPE and MAPE (except Case 2, when compared to MAPE). Second, measures with different error structure motivation seem to perform better when they correspond to the true error structure. Third, Theil's U2, RSE, IQR and KL-divergence based measures perform relatively well. Lastly, the table shows that the measures choose the better forecaster more often when using more observations to evaluate the forecasts.

### 5.1.2 Case 8

We consider another case in which the original DGP is white noise, series length is 30.

We compare two forecasts both generated by independent white noise with the same noise level. Our interest is to see whether the measures wrongly claim one forecast is better than the other though they

are actually the same. In each replication we generate 40 series and evaluate the two forecasts with the eighteen measures. Thus for each replication we produce two series of values of measured performance. We test the null hypothesis of that the two forecasts perform equally well (poor) by a paired  $t$ -test with significance level of 0.05. The empirical size is recorded as the number of rejection of the null based on the accuracy measures. We make 10000 replications and present the mean of the empirical sizes of the test for the 16 measures in Table 4 with different number of periods ( $m = 2, 5, 10$ ). Note that Armstrong and Fildes (1995) suggested that geometric mean might be better than arithmetic mean when evaluating forecasts with multi-series. We introduce the geometric mean of NMSE, Theil's U2 and RAE as GmNMSE, GmTheil'sU2 and GmRAE. We have not observed consistent improvements over the arithmetic mean in our simulation.

From the table, clearly, in general, large  $m$  yields size closer to 0.05 for the measures. MAPE, RSE, and MSEr3 are too conservative. The other measures are satisfactory for this aspect.

### 5.1.3 Case 9

We construct another setting to study the performance of the accuracy measures dealing with series of different natures.

For each replication, we have  $k$  series with series length  $n = 50$ . The data generating process is:  $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + e$ , where for 50 percent of the replications,  $\alpha_0, \alpha_1, \alpha_2, \alpha_3$  are generated as random draw from uniform distribution  $unif(-1, 1)$  and from  $unif(-10, 10)$  for the other half. Here  $x_1, x_2, x_3$  are exogenous variables independently distributed as  $N(0, 1)$  and  $e$  is independent normal distributed as  $N(0, \sigma^2)$  (or double exponential  $DE(0, \sqrt{2}\sigma/2)$ )<sup>3</sup> with  $\sigma = 1$  for 10% of the series and  $\sigma = 0.05$  for the remaining 90%. This way, the different series are not homogenous. We compare the two forecasts  $y^1$  and  $y^2$ , which are generated by:

$$\begin{aligned} y_t^1 &= \alpha_0 + \alpha_1 x_{1t} + \alpha_2 x_{2t} + \alpha_3 x_{3t} \\ y_t^2 &= \hat{\alpha}_{0t} + \hat{\alpha}_{1t} x_{1t} + \hat{\alpha}_{2t} x_{2t} + \hat{\alpha}_{3t} x_{3t} \end{aligned}$$

for  $t = n - m + 1, \dots, n$ , where  $\hat{\alpha}_{0t}, \hat{\alpha}_{1t}, \hat{\alpha}_{2t}, \hat{\alpha}_{3t}$  are estimated by regressing  $y$  on  $x_1, x_2, x_3$ , and a constant term using previous data, i.e.,  $y_1, y_2, \dots, y_{t-1}; x_{1,1}, x_{1,2}, \dots, x_{1,t-1}; x_{2,1}, x_{2,2}, \dots, x_{2,t-1}; x_{3,1}, x_{3,2}, \dots, x_{3,t-1}$ .

For each replication, we sum the numbers produced by the accuracy measures across the 100 series. We declare that a measure chooses the right forecast if the sum of the measured value of  $y_t^1$  is less than that of  $y_t^2$ .

We repeat the replication for 10000 times and record the percentages of choosing the better forecast by the accuracy measures over the replications in Table 5. We also evaluate the three geometric mean methods along with others.

---

<sup>3</sup>We multiply  $\sqrt{2}/2$  to make the variance of the exponential component equal to that of the normal error component. This makes the simulation "fair".

The table suggests that: first, it is better when the number of series used in each replication is larger, which supports the idea of M3-competition that including more series can reduce the influence of dominating series; second, evaluating with five periods is better than evaluating with just two periods; third, geometric means slightly improves for the case of NMSE but not exactly so for Theil's U2 and RAE. MAPE, sMAPE, RSE, and MSEr3 perform poorly relative to others.

## 5.2 Stability of the Accuracy Measures

Stability of accuracy measures is another issue worthy of serious consideration. Since the observations are typically subject to errors, measures that are robust to minor contaminations have an advantage of reliably capturing the performance of the forecasts. With a minor contamination at a sensible level, the more a measure changes, the less it is credible. Obviously, being stable does not qualify an accuracy measure to be a good one, but being unstable with a minor contamination at a level typically seen in an application is definitely a serious weakness.

### 5.2.1 Stability to Linear Transformation

As Clements and Hendry (1993) suggested, stability of accuracy measures with respect to the linear transformation of the original series is an important factor. Here we use a series of monthly Austria/U.S. foreign exchange rate from January 1998 to December 2001. The original series is measured as how many Austrian Schillings are equivalent to one U.S. Dollar. The data was obtained from the web page of Federal Reserve Bank of St. Louis. It is calculated as the average of daily noon buying rates in New York City for cable transfers payable in foreign currencies. We round it to the first digit after the decimal point and perform a linear transformation of the original series by minus the mean of the series and multiply 10, i.e.,

$$y^{new} = 10 \cdot y^{original} - 10 \cdot \text{mean}(y^{original})$$

We have four forecasts generated by random walk, ARIMA(1,1,0), ARIMA(0,1,1), and a forecast generated by a model selected based on BIC criterion from ARIMA models with AR, MA and difference orders from zero to one. Table 6 presents the change of the values produced by the accuracy measures using the last 20 points. We note that the first five accuracy measures produced very different values after the transformation since they are not location-scale transformation invariant. Note also that the last three accuracy measures had some minor changes. This suggested that the first five measures are generally not good for cross-series comparison of forecasting procedures since a linear transformation of the original series may change the ranking of the forecast.

### 5.2.2 Stability to Perturbation

In addition to robustness to linear transformation, a good accuracy measure should be robust to measurement error. It is common that available quantities are subject to some disturbances, e.g., due to

rounding, truncation or measurement errors. When the original series ( $F$ ) is added with a disturbance term simulating the rounded digit, the accuracy measures may produce a different ranking of the forecasts. The change of the best ranked forecast indicates the instability of the accuracy measures with respect to such a disturbance. In addition, we can add a small normally distributed disturbance on the original series.

The data set used is Earnings Yield of All Common Stocks on the New York Stock Exchange from 1871 to 1918. The series is obtained from NBER (National Bureau of Economic Research) website. The unit is percent and the numbers are rounded to two decimals. We have two forecasts: one generated by random walk and the other from ARIMA with AR, MA and difference orders selected by BIC over the choices of zero and one. The forecasts are ranked using the accuracy measures. We perturb the data by adding a small disturbance.

(1) Rounding:  $F' = F + u$ , where  $u$  is generated from a uniform distribution  $Unif(-0.005, 0.005)$ . This addition is used to simulate the actual numbers which were rounded up to two decimals (as given in the data).

(2) Truncation:  $F' = F + u$ , where  $u$  is generated from a uniform distribution  $Unif(0, 0.01)$ . This is used to simulate the actual numbers assuming that the numbers in the data were truncated up to two decimals.

(3) Normal 1:  $F' = F + e$ , where  $e$  is random draw from a normal distribution of  $N(0, (0.1\sigma_F)^2)$ , where  $\sigma_F$  is the sample standard deviation of the original series.

(4) Normal 2:  $F' = F + e$ , where  $e$  is random draw from  $N(0, (0.12\sigma_F)^2)$ .

The perturbation is replicated 5000 times. Then we can make forecasts based on the perturbed dataset and obtain the new ranking of the two different forecast methods. Table 7 shows the percentage change for the earnings yield data set. Note that  $KL - N1$ ,  $KL - N2$ ,  $MSEr1$ ,  $MSEr2$ ,  $MSEr3$  are relatively unstable when subject to rounding, truncation, or normal perturbation. The poor performance of these measures is probably due to the poor variance estimation in the denominator of the measures. It is rather surprising that  $RSE$  performs so well in this example, but we suspect that this does not hold generally. Note that  $RSE$  faces a problem when the denominator happens to be close to zero, which is reflected in its poor performance shown in the earlier tables. Its modification  $mRSE$  addresses this difficulty and has a good overall performance. Not surprisingly, the measures are less stable when the variance of the normal perturbation is greater. Even though  $MAPE$  performs well under rounding and normal perturbations, it is highly unstable when truncation is involved.

### 5.3 Evaluating at one point vs. evaluating at multiple points

As to how many points we should use to compare different forecasts under MSE based on a single series, Ashley (2003) presented an analysis from statistical significance point of view. For cross-series comparison, our earlier experiments suggest that the ability of choosing the better forecast improves significantly when using more points for the evaluation as found in Tables 3, 4 and 5. Another observation

is that when  $m$  is small, accuracy measures of different error structure motivation perform more similarly than when  $m$  is large. An extreme example is that linear loss function and absolute value loss function are equivalent when  $m = 1$ .

## 6 Concluding Remarks

In this paper, we studied various forecast accuracy measures. Theoretically speaking, for comparing two forecasters, only when the errors are stochastically ordered, the ranking of the forecasts is basically independent of the form of the chosen accuracy measure. Otherwise, the ranking depends on the specification of the accuracy measure. Under some conditions on the conditional distribution of  $Y$ , K-L divergence based accuracy measures are well-motivated and have certain nice invariance properties.

In the empirical direction, we studied the performance of the familiar accuracy measures and some new ones. They were compared in two important aspects: in selecting the known-to-be-better forecaster and the robustness when subject to random disturbance, e.g., measurement error.

The results suggest the following:

(1) For cross-series comparison of forecasts, individually tailored measures may improve the performance of differentiating between good and poor forecasters. More work needs to be done on how to select a measure based on the characteristics of each individual series. For example, we may use a QQ plot and/or other means to have a good sense on the shape of the error distribution and then apply the corresponding accuracy measures.

(2) Stochastically ordered forecast errors provide a tool for objectively comparing different forecast accuracy measures by assessing their ability to choose the better or best forecast.

(3) In addition to the known facts that MAPE and sMAPE are not location invariant, and they have a major flaw when the true value of the forecast is close to zero, we obtained new information on MAPE and related measures: their ability to pick out the better forecast is substantially worse than the other accuracy measures. The proposed msMAPE showed a significant improvement over MAPE and sMAPE in this aspect. The MSE based relative measures are generally better than MAPE and sMAPE, but not as good as K-L divergence based measures.

(4) We proposed the well motivated KL-divergence and IQR based measures, which were shown to have relatively good performance in the simulations.

## 7 Acknowledgments

The work of the second author was supported by the United States National Science Foundation CAREER Award Grant DMS-00-94323.

## References

- [1] Ahlburg, A (1992) "A commentary on error measures: Error measures and the choice of a forecast method", *International Journal of Forecasting*, Vol 8, pp99-100

- [2] Armstrong, S. & R. Fildes (1995), "On the Selection of Error Measures for Comparisons Among Forecasting Methods," *Journal of Forecasting*, 14, 67-71
- [3] Ashley, R. "Statistically significant forecasting improvements: how much out-of-sample data is likely necessary?", *International Journal of Forecasting*, Volume 19, Issue 2, April-June 2003, Pages 229-239
- [4] Barron, A.R. (1987). "Are Bayes rules consistent in information?" *Open Problems in Communication and Computation*, pp85-91. T.M. Cover and B. Gopinath eds., Springer, NY.
- [5] Carbone R. and J.S. Armstrong, 1982, "Evaluation of extrapolative forecasting methods: Results of a survey of academicians and practitioners," *Journal of Forecasting* 1, 215-217.
- [6] Christoffersen, P. F. and F.X. Diebold (1998), "Cointegration and Long Horizon Forecasting," *Journal of Business and Economic Statistics*, v. 15.
- [7] Clements, M.P. and D.F. Hendry (1993). "On the limitations of comparing mean squared forecast errors". *Journal of Forecasting*, 12, 617-637. With discussion
- [8] Collopy, F., & Armstrong, J. S. (1992a). "Rule-based forecasting". *Management Science* 38, 1394-1414
- [9] Collopy, F., & Armstrong, J. S. (1992b) "Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons", *International Journal of Forecasting*, 8 (1992), 69-80
- [10] Cover, T.M., & J.A. Thomas *Elements of Information Theory*, 1991, John Wiley and Sons.
- [11] Diebold, F.X.; R. Mariano (1995) "Comparing forecast accuracy", *Journal of Business Economics and Statistics*, Vol 13, pp253-265
- [12] Goodwin, P. & Lawton, R. (1999) "On the asymmetry of the symmetric MAPE", *International Journal of Forecasting*, Vol. 15, No.4, pp405-408
- [13] Granger, C. W. J.; Jeon, Y. (2003) "A Time-Distance Criterion for Evaluating forecasting models", *International Journal of Forecasting*, Vol 19, pp199-215
- [14] Granger, C. W. J.; Jeon, Y. (2003) "Comparing Forecasts of Inflation Using Time Distance", *International Journal of Forecasting*, Vol 19, pp339-349
- [15] Granger, C. W. J.; Pesaran, M. H. (2000) "Economic and Statistical Measures of Forecast Accuracy", *Journal of Forecasting*, Vol 19, pp537-560
- [16] Koehler, A.B. (2001) "The asymmetry of the sAPE measure and other comments on the M3-competition", *International Journal of Forecasting*, Vol. 17, pp570-574
- [17] Makridakis, S. (1993) "Accuracy measures: theoretical and practical concerns", *International Journal of Forecasting*, Vol.9, pp527-529
- [18] Makridakis, S.; Hibon, M., & Moser, C. (1979) "Accuracy of Forecasting: An Empirical Investigation", *Journal of the Royal Statistical Society, Series A (General)* Vol. 142, Issue 2, pp97-145
- [19] Makridakis, S.; & Hibon, M. (2000) "The M3-Competition: results, conclusions and implications", *International journal of Forecasting* Vol. 16, pp451-476
- [20] Newbold, P., "The competition to end all competitions," *Journal of Forecasting*, 2 (1983), 276-9.
- [21] Tashman, L. J. (2000) "Out-of-sample tests of forecasting accuracy: an analysis and review", *International Journal of Forecasting*, Vol. 16, pp437-450
- [22] Thompson, P.A. (1990) "An MSE statistic for comparing forecast accuracy across series", *International Journal of Forecasting*, Vol. 6, pp219-227
- [23] Yang, Y. and Barron, A.R. (1999) "Information-theoretic determination of minimax rates of convergence", *Ann. Statistics*, 27, 1564-1599.
- [24] Yang, Y. (2000) "Mixing Strategies for Density Estimation", *Annals of Statistics* , vol. 28, pp. 75-87.
- [25] Yokum, J.T., and Armstrong J. S. (1995) "Beyond Accuracy: Comparison of Criteria Used to Select Forecasting Methods", *International Journal of Forecasting*, Vol. 11, pp591-597

|                  | Case 1 |       | Case 2 |       | Case 3 |       | Case 4 |       | Case 5 |       | Case 6 |       | Case 7 |       |
|------------------|--------|-------|--------|-------|--------|-------|--------|-------|--------|-------|--------|-------|--------|-------|
| $m$              | 20     | 2     | 20     | 2     | 20     | 2     | 20     | 2     | 20     | 2     | 20     | 2     | 20     | 2     |
| <i>MAPE</i>      | 0.596  | 0.546 | 0.999  | 0.764 | 0.830  | 0.750 | 0.639  | 0.568 | 0.700  | 0.594 | 0.786  | 0.648 | 0.736  | 0.617 |
| <i>sMAPE</i>     | 0.623  | 0.506 | 0.000  | 0.000 | 0.963  | 0.729 | 0.675  | 0.560 | 0.760  | 0.588 | 0.751  | 0.578 | 0.726  | 0.572 |
| <i>msMAPE</i>    | 0.688  | 0.530 | 0.575  | 0.537 | 0.986  | 0.771 | 0.739  | 0.574 | 0.823  | 0.600 | 0.815  | 0.614 | 0.780  | 0.599 |
| <i>RMSE</i>      | 0.757  | 0.544 | 0.979  | 0.721 | 0.996  | 0.791 | 0.825  | 0.576 | 0.824  | 0.592 | 0.938  | 0.662 | 0.841  | 0.627 |
| <i>NMSE</i>      | 0.757  | 0.544 | 0.979  | 0.721 | 0.996  | 0.791 | 0.825  | 0.576 | 0.824  | 0.592 | 0.938  | 0.662 | 0.841  | 0.627 |
| <i>KL - N</i>    | 0.760  | 0.544 | 0.977  | 0.720 | 0.996  | 0.791 | 0.824  | 0.576 | 0.821  | 0.592 | 0.937  | 0.662 | 0.840  | 0.627 |
| <i>KL - N1</i>   | 0.709  | 0.546 | 0.946  | 0.711 | 0.979  | 0.777 | 0.778  | 0.576 | 0.770  | 0.593 | 0.910  | 0.661 | 0.823  | 0.626 |
| <i>KL - N2</i>   | 0.752  | 0.547 | 0.966  | 0.717 | 0.992  | 0.788 | 0.807  | 0.577 | 0.799  | 0.591 | 0.931  | 0.662 | 0.836  | 0.625 |
| <i>KL - DE1</i>  | 0.758  | 0.543 | 0.976  | 0.721 | 0.996  | 0.790 | 0.819  | 0.580 | 0.840  | 0.596 | 0.929  | 0.661 | 0.860  | 0.626 |
| <i>KL - DE2</i>  | 0.757  | 0.543 | 0.975  | 0.720 | 0.996  | 0.789 | 0.817  | 0.579 | 0.844  | 0.596 | 0.928  | 0.661 | 0.860  | 0.625 |
| <i>IQR</i>       | 0.758  | 0.544 | 0.977  | 0.720 | 0.996  | 0.791 | 0.822  | 0.577 | 0.820  | 0.593 | 0.934  | 0.663 | 0.841  | 0.627 |
| <i>RSE</i>       | 0.633  | 0.541 | 0.836  | 0.701 | 0.992  | 0.851 | 0.600  | 0.564 | 0.642  | 0.581 | 0.699  | 0.639 | 0.671  | 0.613 |
| <i>mRSE</i>      | 0.781  | 0.549 | 0.986  | 0.721 | 1.000  | 0.824 | 0.763  | 0.574 | 0.792  | 0.589 | 0.911  | 0.656 | 0.817  | 0.623 |
| <i>Theil'sU2</i> | 0.757  | 0.544 | 0.979  | 0.721 | 0.996  | 0.791 | 0.825  | 0.576 | 0.824  | 0.592 | 0.938  | 0.662 | 0.841  | 0.627 |
| <i>RAE</i>       | 0.724  | 0.544 | 0.970  | 0.716 | 0.995  | 0.788 | 0.784  | 0.577 | 0.854  | 0.602 | 0.925  | 0.659 | 0.861  | 0.624 |
| <i>MSEr1</i>     | 0.610  | 0.521 | 0.916  | 0.674 | 0.975  | 0.747 | 0.658  | 0.557 | 0.776  | 0.591 | 0.864  | 0.636 | 0.810  | 0.615 |
| <i>MSEr2</i>     | 0.691  | 0.529 | 0.953  | 0.647 | 0.984  | 0.708 | 0.759  | 0.550 | 0.740  | 0.571 | 0.902  | 0.610 | 0.796  | 0.589 |
| <i>MSEr3</i>     | 0.686  | 0.529 | 0.926  | 0.647 | 0.961  | 0.708 | 0.749  | 0.550 | 0.730  | 0.571 | 0.865  | 0.610 | 0.778  | 0.589 |

Table 3: Percentage of Choosing the best model

|                    | $m = 2$ | $m = 5$ | $m = 10$ |
|--------------------|---------|---------|----------|
| <i>MAPE</i>        | 0.022   | 0.020   | 0.021    |
| <i>sMAPE</i>       | 0.057   | 0.045   | 0.051    |
| <i>msMAPE</i>      | 0.057   | 0.047   | 0.052    |
| <i>RMSE</i>        | 0.053   | 0.052   | 0.050    |
| <i>NMSE</i>        | 0.044   | 0.048   | 0.049    |
| <i>KL - N</i>      | 0.051   | 0.051   | 0.051    |
| <i>KL - N1</i>     | 0.051   | 0.051   | 0.048    |
| <i>KL - N2</i>     | 0.053   | 0.050   | 0.049    |
| <i>KL - DE1</i>    | 0.050   | 0.049   | 0.051    |
| <i>KL - DE2</i>    | 0.051   | 0.049   | 0.050    |
| <i>IQR</i>         | 0.052   | 0.051   | 0.051    |
| <i>RSE</i>         | 0.018   | 0.021   | 0.018    |
| <i>mRSE</i>        | 0.051   | 0.050   | 0.054    |
| <i>Theil'sU2</i>   | 0.039   | 0.050   | 0.050    |
| <i>RAE</i>         | 0.040   | 0.047   | 0.048    |
| <i>GmNMSE</i>      | 0.055   | 0.050   | 0.049    |
| <i>GmTheil'sU2</i> | 0.055   | 0.050   | 0.049    |
| <i>GmRAE</i>       | 0.055   | 0.051   | 0.050    |
| <i>MSEr1</i>       | 0.053   | 0.050   | 0.050    |
| <i>MSEr2</i>       | 0.045   | 0.041   | 0.041    |
| <i>MSEr3</i>       | 0.024   | 0.020   | 0.018    |

Table 4: Empirical Size of the Paired  $t$  test

| # of series        | 60           |       |             |       | 20           |       |             |       |
|--------------------|--------------|-------|-------------|-------|--------------|-------|-------------|-------|
| Error              | Normal Error |       | Double Exp. |       | Normal Error |       | Double Exp. |       |
| $m$                | 5            | 2     | 5           | 2     | 5            | 2     | 5           | 2     |
| <i>MAPE</i>        | 0.750        | 0.736 | 0.837       | 0.792 | 0.735        | 0.663 | 0.792       | 0.794 |
| <i>sMAPE</i>       | 0.776        | 0.749 | 0.863       | 0.814 | 0.751        | 0.678 | 0.802       | 0.793 |
| <i>msMAPE</i>      | 0.997        | 0.939 | 0.999       | 0.978 | 0.925        | 0.818 | 0.973       | 0.902 |
| <i>RMSE</i>        | 1.000        | 0.981 | 1.000       | 0.993 | 0.979        | 0.871 | 0.990       | 0.928 |
| <i>NMSE</i>        | 0.997        | 0.892 | 0.998       | 0.932 | 0.950        | 0.786 | 0.966       | 0.853 |
| <i>KL - N</i>      | 1.000        | 0.964 | 0.999       | 0.986 | 0.970        | 0.863 | 0.980       | 0.906 |
| <i>KL - N1</i>     | 0.998        | 0.954 | 0.998       | 0.987 | 0.961        | 0.847 | 0.972       | 0.890 |
| <i>KL - N2</i>     | 1.000        | 0.956 | 0.998       | 0.982 | 0.972        | 0.862 | 0.981       | 0.904 |
| <i>KL - DE1</i>    | 0.973        | 0.906 | 0.975       | 0.908 | 0.941        | 0.838 | 0.939       | 0.842 |
| <i>KL - DE2</i>    | 0.973        | 0.909 | 0.971       | 0.909 | 0.939        | 0.836 | 0.941       | 0.847 |
| <i>IQR</i>         | 1.000        | 0.963 | 0.999       | 0.986 | 0.967        | 0.849 | 0.980       | 0.903 |
| <i>RSE</i>         | 0.701        | 0.707 | 0.772       | 0.771 | 0.687        | 0.661 | 0.721       | 0.730 |
| <i>mRSE</i>        | 0.997        | 0.948 | 0.997       | 0.980 | 0.954        | 0.850 | 0.968       | 0.887 |
| <i>Theil'sU2</i>   | 0.999        | 0.913 | 0.998       | 0.941 | 0.961        | 0.815 | 0.973       | 0.860 |
| <i>RAE</i>         | 0.993        | 0.891 | 0.998       | 0.952 | 0.937        | 0.797 | 0.974       | 0.879 |
| <i>GmNMSE</i>      | 0.999        | 0.907 | 1.000       | 0.979 | 0.965        | 0.791 | 0.977       | 0.899 |
| <i>GmTheil'sU2</i> | 0.999        | 0.907 | 1.000       | 0.979 | 0.965        | 0.791 | 0.977       | 0.899 |
| <i>GmRAE</i>       | 0.998        | 0.898 | 1.000       | 0.986 | 0.950        | 0.788 | 0.987       | 0.909 |
| <i>MSEr1</i>       | 0.972        | 0.862 | 0.999       | 0.987 | 0.884        | 0.762 | 0.968       | 0.899 |
| <i>MSEr2</i>       | 0.936        | 0.827 | 0.949       | 0.879 | 0.858        | 0.715 | 0.875       | 0.791 |
| <i>MSEr3</i>       | 0.782        | 0.720 | 0.823       | 0.757 | 0.762        | 0.665 | 0.796       | 0.710 |

Table 5: Percentage of Choosing the Better Forecaster

| forecast         | Random Walk |       | ARIMA(1,1,0) |       | ARIMA(0,1,1) |       | ARIMA (BIC) |       |
|------------------|-------------|-------|--------------|-------|--------------|-------|-------------|-------|
| series           | original    | new   | original     | new   | original     | new   | original    | new   |
| <i>MAPE</i>      | 0.024       | 0.302 | 0.023        | 0.306 | 0.022        | 0.296 | 0.030       | 0.381 |
| <i>sMAPE</i>     | 0.024       | 0.280 | 0.023        | 0.264 | 0.022        | 0.256 | 0.030       | 0.371 |
| <i>msMAPE</i>    | 0.023       | 0.207 | 0.022        | 0.196 | 0.021        | 0.190 | 0.029       | 0.263 |
| <i>RMSE</i>      | 0.428       | 4.278 | 0.431        | 4.305 | 0.421        | 4.207 | 0.569       | 5.489 |
| <i>NMSE</i>      | 0.426       | 0.135 | 0.428        | 0.135 | 0.423        | 0.134 | 0.492       | 0.153 |
| <i>KL - N</i>    | 0.410       | 0.410 | 0.415        | 0.415 | 0.409        | 0.409 | 0.590       | 0.574 |
| <i>KL - N1</i>   | 0.869       | 0.869 | 0.822        | 0.822 | 0.805        | 0.805 | 1.147       | 1.089 |
| <i>KL - N2</i>   | 0.677       | 0.677 | 0.666        | 0.666 | 0.649        | 0.649 | 0.853       | 0.830 |
| <i>KL - DE1</i>  | 0.071       | 0.071 | 0.071        | 0.071 | 0.069        | 0.069 | 0.132       | 0.122 |
| <i>KL - DE2</i>  | 0.109       | 0.109 | 0.109        | 0.109 | 0.106        | 0.106 | 0.202       | 0.188 |
| <i>IQR</i>       | 0.398       | 0.398 | 0.419        | 0.419 | 0.414        | 0.414 | 0.626       | 0.611 |
| <i>RSE</i>       | 0.975       | 0.975 | 1.158        | 1.158 | 1.113        | 1.113 | 1.612       | 1.470 |
| <i>mRSE</i>      | 0.348       | 0.348 | 0.347        | 0.347 | 0.341        | 0.341 | 0.501       | 0.478 |
| <i>Theil'sU2</i> | 1.000       | 1.000 | 1.006        | 1.006 | 0.983        | 0.983 | 1.331       | 1.283 |
| <i>RAE</i>       | 1.000       | 1.000 | 0.965        | 0.965 | 0.934        | 0.934 | 1.247       | 1.156 |
| <i>MSEr1</i>     | 1.078       | 1.100 | 0.929        | 0.934 | 0.870        | 0.878 | 1.104       | 1.071 |
| <i>MSEr2</i>     | 0.703       | 0.710 | 0.726        | 0.733 | 0.705        | 0.711 | 0.880       | 0.864 |
| <i>MSEr3</i>     | 0.729       | 0.739 | 0.752        | 0.762 | 0.731        | 0.740 | 0.912       | 0.899 |

Table 6: Stability of Accuracy Measure to Linear Transformation

|                  | Rounding | Truncation | Normal 1 | Normal 2 |
|------------------|----------|------------|----------|----------|
| <i>MAPE</i>      | 0.068    | 0.959      | 0.213    | 0.520    |
| <i>sMAPE</i>     | 0.092    | 0.024      | 0.362    | 0.704    |
| <i>msMAPE</i>    | 0.089    | 0.025      | 0.431    | 0.728    |
| <i>RMSE</i>      | 0.056    | 0.043      | 0.620    | 0.877    |
| <i>NMSE</i>      | 0.056    | 0.043      | 0.619    | 0.877    |
| <i>KL - N</i>    | 0.089    | 0.070      | 0.612    | 0.867    |
| <i>KL - N1</i>   | 0.278    | 0.896      | 0.368    | 0.134    |
| <i>KL - N2</i>   | 0.148    | 0.062      | 0.598    | 0.861    |
| <i>KL - DE1</i>  | 0.061    | 0.030      | 0.605    | 0.831    |
| <i>KL - DE2</i>  | 0.047    | 0.024      | 0.602    | 0.828    |
| <i>IQR</i>       | 0.071    | 0.053      | 0.611    | 0.874    |
| <i>RSE</i>       | 0.004    | 0.001      | 0.017    | 0.067    |
| <i>mRSE</i>      | 0.018    | 0.010      | 0.249    | 0.356    |
| <i>Theil'sU2</i> | 0.056    | 0.043      | 0.620    | 0.877    |
| <i>RAE</i>       | 0.044    | 0.031      | 0.541    | 0.873    |
| <i>MSEr1</i>     | 0.169    | 0.859      | 0.604    | 0.170    |
| <i>MSEr2</i>     | 0.201    | 0.077      | 0.571    | 0.634    |
| <i>MSEr3</i>     | 0.243    | 0.086      | 0.562    | 0.627    |

Table 7: Rate of ranking change