A statistical test for forecast evaluation under a discrete loss function

Francisco J. Eransus, Alfonso Novales
Departamento de Economia Cuantitativa
Universidad Complutense

December 2010

Abstract

We propose a new approach to evaluating the usefulness of a set of forecasts, based on the use of a discrete loss function defined on the space of data and forecasts. Existing procedures for such an evaluation either do not allow for formal testing, or use tests statistics based just on the frequency distribution of (data , forecasts)-pairs. They can easily lead to misleading conclusions in some reasonable situations, because of the way they formalize the underlying null hypothesis that '*the set of forecasts is not useful*'. Even though the ambiguity of the underlying null hypothesis precludes us from performing a standard analysis of the size and power of the tests, we get results suggesting that the proposed DISC test performs better than its competitors.

Keywords: Forecasting Evaluation, Loss Function.
JEL Classification:

1 Introduction

To evaluate whether a set of forecasts is acceptable, a global distance between actual data and the associated forecasts is generally computed, usually through a smooth continuous function of the forecast errors like the mean absolute error or the mean squared error. While this may be convenient to establish some comparison between alternative models, it is not so useful to judge the quality of a single set of forecasts. The alternative approach proceeds to formally testing for the quality of forecasts. A first class of tests is based on a standard measure of association between data and forecasts, that could take the form of a significance test on the slope of a regression of realized data on the previously obtained forecasts. A second class of tests is based on a two-way classification of $(data, forecast)$-pairs on the basis of a finite partition of the data space. These tests consider the ambiguous null hypothesis that "the set of forecasts is not useful" and identify lack of usefulness with some characteristic of the joint frequency distribution of data and forecasts along the lines of them being independent from each other. This can be implemented either by a Pearson chi-square test on a contingency table (CT) or by the popular Pesaran and Timmermann (1992) (PT) test. As a special case, the literarure on macroeconomic forecasting has sometimes used the proposal by Merton (1981) and Henriksson and Merton (1981) to consider a two-region partition of the data space to test for whether forecasts and actual data fall at either side of a given numerical reference in an independent manner. When such reference is zero, as in the binomial test

proposed in Greer (2003), the test examines the independence of signs in forecasts and actual data.[1]

Unfortunately, neither approach is completely satisfactory. All the mentioned tests share two important limitations: *a)* while being appropriate for qualitative data, they do not fully exploit the information contained in the numerical values of each data point and its associated forecast. The size of the forecast error does not play any specific role in any of these tests and, with the exception of those tests using a two-region partition, the tests do not specifically take into account whether the prediction of the sign was correct. This is too restrictive for most situations in Economics and Finance, in which a detailed comparison of data and forecasts is needed, *b)* the definition of forecast quality in all the mentioned tests is independent of the forecasting context, with the user not playing any role in specifying that definition. This situation is again not too reasonable, since the cost of missing the sign of the data or the implications of making a given forecast error is bound to differ for each particular forecasting application.

This paper alleviates these restraints by introducing a new approach through a formal test for "usefulness" of a set of forecasts, that we will denote by DISC. Like the CT and PT tests, the DISC test is based on a two-way classification of the $(data, forecast)$-pairs. The relevant difference is that the DISC test is based on a loss function defined on the classification table and compares the observed mean loss with the expected loss under independence of data and forecasts. Hence, at a difference of previous tests, the DISC test analyses the frequency distribution of the loss function, rather than the bidimensional distribution of data and forecasts.

Assigning weights to each cell in the classification table through the loss function, the user incorporates the costs associated to each $(data, forecast)$-pair for each particular setup, thereby solving the limitations described in *a)* and *b)*. The need to quantify before hand the cost of each $(data, forecast)$-pair, with the results of the test being conditional on such characterization, should be seen as a strength of our proposal, rather than as a weakness. The alternative of specifying a continuous loss function like the squared forecast error or its absolute value evades this issue by imposing a very tight structure on the loss function without considering whether such structure is really appropriate for the forecasting application in hand, or without assessing how does it condition the result of the forecasting evaluation exercise. That way, our comments *a)* and *b)* above apply in full force.

Discrete loss functions allow for the incorporation of many types of asymmetries and nonlinearities. By placing an upper bound on the value of the loss, they also limit the potential influence of an occasionally large forecast error that could condition the result of the forecast evaluation. Furthermore, there are situations in which they are necessary, as it is the case when dealing with qualitative data. The discrete setup has also a technical advantage, since it allows us to readily characterize the asymptotic distribution of our proposed test statistic.

We compare the DISC test with its natural competitors, CT and PT, and our results suggest that the former may perform better than the latter in many interesting and realistic setups. The DISC test is more powerful in situations where the set of forecasts are evidently useful. In cases where there may be some doubt about the usefulness of forecasts, the behavior of the DISC test is more reasonable, in the sense that the decision reached will depend in a natural way on the specific loss function chosen by the user.

The DISC test falls in the category of tests evaluating whether a set of forecasts is useful

[1] Practical aplications to macroeconomic or financial forecasts of the last two types of tests can be found in Schnader and Stekler (1990), Stekler (1994), Leitch and Tanner (1995), Kolb and Stekler (1996), Ash, Smyth an Heravi (1998), Mills and Pepper (1999), Joutz and Stekler (2000), Oller and Bharat (2000), Pons (2000), Greer (2003) among many others.

or acceptable. When the result of such evaluation is positive, the set of forecasts should be evaluated for rationality or efficiency along the lines described in Joutz and Stekler (2000), by testing for a zero drift and unit slope in the linear projection of data on forecasts, lack of serial correlation in the residuals and zero correlation between residuals and any information that was available at the time the forecasts were made.

The paper is organized as follows: in Section 2 we discuss the difficulties associated with the CT and PT tests. In Section 3 we introduce our proposal: we explain the general approach, describe discrete loss functions and derive the DISC test. In Section 4 we analyze the performance of the DISC test relative to the CT and PT tests through simulation exercises. Finally, Section 5 summarizes the main conclusions.

2 Criticism of standard tests

We denote by y_t the time t realization of the variable being forecasted, and by $\widehat{y}_t$ the associated forecast made at $t-h$. We assume that we have a set of T $(y_t, \widehat{y}_t)$-pairs. The CT and PT test divide the data domain in m regions and therefore, the bidimensional domain of data and forecasts is partitioned into m^2-squares. CT is a nonparametric Pearson test with null hypothesis $p_{ij} = p_{i.}p_{.j}$, where p_{ij} is the joint probability that y_t falls in the i-th region while the forecast $\widehat{y}_t$ falls in the j-th region, and $p_{i.}$, $p_{.j}$ denote the marginal probabilities that y_t and $\widehat{y}_t$ fall in the i-th and j-th regions, respectively. The PT test considers the null hypothesis that $\sum_{i=1}^{m} p_{ii} = \sum_{i=1}^{m} p_{i.}p_{.i}$, and uses the natural statistic $\sum_{i=1}^{m} \widehat{p}_{ii} - \sum_{i=1}^{m} \widehat{p}_{i.}\widehat{p}_{.i}$ that substitutes relative sample frequencies for probabilities, which follows a $N(0,1)$ distribution after normalizing by the corresponding standard deviation [see Pesaran and Timmermann (1992)]. PT is usually implemented as a one-side test, taking just the upper-tail of the $N(0,1)$ distribution as the critical region. CT is a non parametric test of stochastic independence, while the PT test is less restrictive, since it just evaluates the quadrants along the main diagonal of the bidimensional table of frequencies. According to Pesaran and Timmermann (1992), the CT test is, in general, more conservative than the PT test.

As mentioned in the Introduction, the use of these tests to evaluate the predictive ability of a single model is not fully appropriate, since a possible rejection of the null hypothesis is not too informative about forecast quality, as the following example illustrates. Let us assume that we use a partition of the space of data and forecasts into four regions: L-, S-, S+, L+, where the '+', '-' signs indicate the sign of the data/forecast, while 'L', 'S' denote whether the data/forecast are large or small in absolute value, this difference defined by certain threshold. Suppose we have a sample of 100 data points on the period-to-period rate of change of a given time series and the associated forecasts obtained from three alternative forecasting models. The hypothetical information on the predictive results has been summarized in the matrices that follow, whose elements represent the absolute frequencies observed in each cell of the joint partition of data and forecasts:[2]

$$M_1 = \begin{array}{cc} & \\ & y_t \end{array} \begin{array}{c} \\ \text{L-} \\ \text{S-} \\ \text{S+} \\ \text{L+} \end{array} \begin{array}{c} \widehat{y}_t \\ \text{L-} \quad \text{S-} \quad \text{S+} \quad \text{L+} \\ \hline \mathbf{0} \quad \mathbf{0} \quad 0 \quad \textit{25} \\ \mathbf{0} \quad \mathbf{0} \quad 25 \quad \textit{0} \\ \textit{0} \quad 25 \quad \mathbf{0} \quad \mathbf{0} \\ \textit{25} \quad 0 \quad \mathbf{0} \quad \mathbf{0} \end{array}$$

[2] Actually, the test results we mention were obtained after changing the single 25 value that appears in the first row of each matrix by 24, while the (1,3)-element is 1, rather than 0. This was done to avoid the variance of the PT statistic to be zero.

$$M_2 = \begin{array}{cc|cccc} & & & & \widehat{y}_t & \\ & & \text{L-} & \text{S-} & \text{S+} & \text{L+} \\ \hline & \text{L-} & \mathbf{0} & \mathbf{25} & 0 & \textit{0} \\ y_t & \text{S-} & \mathbf{25} & \mathbf{0} & 0 & \textit{0} \\ & \text{S+} & \textit{0} & 0 & \mathbf{0} & \mathbf{25} \\ & \text{L+} & \textit{0} & 0 & \mathbf{25} & \mathbf{0} \end{array} \qquad M_3 = \begin{array}{cc|cccc} & & & & \widehat{y}_t & \\ & & \text{L-} & \text{S-} & \text{S+} & \text{L+} \\ \hline & \text{L-} & \mathbf{25} & \mathbf{0} & 0 & \textit{0} \\ y_t & \text{S-} & \mathbf{0} & \mathbf{25} & 0 & \textit{0} \\ & \text{S+} & \textit{0} & 0 & \mathbf{25} & \mathbf{0} \\ & \text{L+} & \textit{0} & 0 & \mathbf{0} & \mathbf{25} \end{array}$$

In most applications it would be desirable that the forecasts had the right sign and be as precise as possible, in the sense that the distance between the region of the partition where the forecast falls was as close as possible to that of the data. In the previous matrices we have indicated in boldface the squares where the forecasts would be correct in sign. In italics we denote those squares where the forecasts would not only have the wrong sign, but also they would have the least precision possible.

Forecasts from Model 1 (M1) have always the wrong sign. They also have the least possible precision 50% of the times, those in cells (1,4) and (4,1). Model 2 (M2) always forecasts the sign correctly, but it never reaches the highest precision, as reflected in an empty main diagonal. Model 3 (M3) is always fully right, in sign as well as in magnitude. Even though forecasts are not stochastically independent from the data for any of the three models, it is evident that the M1 forecasts have no value whatsoever for the user, those from M3 are optimal given the partition of the data space, while those from M2 might be useful, even though they are not fully precise. The reasonable outcome would be that the tests would not reject the lack of utility of M1 forecasts, rejecting it for M3 forecasts. The desired result for M2 forecasts should depend on the specific forecasting context and the specific definition by the user of what is meant by 'useful predictions'.

Unfortunately, the CT test rejects the null hypothesis in the three cases. The PT test rejects the null for M3 and it does not reject it for M1 and M2.[3] Furthermore, the numerical value of the CT test statistic is the same for the three models, 292.31, while that of the PT test statistic is the same for M1 and M2, -353.55. Three situations as different as those in the example are indistinguishable for the CT test, while M1 and M2 are indistinguishable from the point of view of the PT test.

The example illustrates the potential errors made by existing tests of 'lack of usefulness' of a set of forecasts: *i*) for M1, the CT test detects some stochastic dependence between data and forecasts, thereby leading to the rejection of the null hypothesis of independence. This is because the test does not pay attention to the type of dependence, which happens to be negative in this case, *ii*) for M2, the error comes about from the fact that the CT and PT tests do not take into account any characteristic of the forecast error, like its sign or size, whenever the forecast falls in a region different from that of the data. And this is a consequence of both tests using too general an approach to the concept of 'useful forecasts', without considering the possibility that the definition of such concept might depend on the specific forecasting situation. For instance, a model that produces forecasts that are always right in sign but not in magnitude may be very useful in some applications but not so much in some other setups. Not to mention the convenience of taking into account the size of the forecast error. These are some of the limitations we mentioned in the Introduction.

[3] We have used PT as a one-sided test, rejecting the null hypothesis when the test statistic test is positive and large enough.

3 A proposal based on a discrete loss function

We now present an approach alternative to those of the CT and PT tests which, incorporating a specific type of loss functions, may greatly alleviate the limitations of these two tests.

3.1 General overview

Let $f(y_t, \widehat{y}_t)$ be a loss function on two arguments, the data and the associated forecast. We propose testing the hypothesis $H_0 \equiv E(f_t) - E(f_t^{IE}) = 0$ ('forecasts are not useful') against $H_1 \equiv E(f_t) - E(f_t^{IE}) < 0$ ('forecasts are useful'), where f_t^{IE} denotes the loss that would arise with a set of forecasts independent from the data. We consider a set of forecasts not to be useful when the mean loss is at least as large as the one that would obtain from f_t^{IE}. Our approach does not pretend to solve completely the ambiguity in the specification of the null hypothesis when testing for predictive ability, but it is more satisfactory than that of the CT and PT tests. On the one hand, the incorporation of a loss function allows us to define a one-sided alternative hypothesis, avoiding potential mistakes as those made by the CT test under the M1 forecasts. On the other hand, the definition of what is meant by a 'not useful' set of forecasts can be made explicit through the choice of a loss function f . Precisely, under a discrete loss function f , it is easy to adjust that definition to each particular application, as we explain below in some examples. Relative to previous tests, the use of a discrete loss function amounts to assigning a different weight to each $(data, forecast)$-pair. That changes in a non trivial way the frequency distribution of $(data, forecast)$-pairs. Besides, the significance of our proposal can be seen in that our test is no longer based on the frequency distribution of the $(data, forecast)$-pairs but rather, on the implied frequency distribution of the loss function.

After partitioning the domain of y_t and $\widehat{y}_t$ in m regions, so that the joint domain of data and forecasts is naturally partitioned in m^2 cells, we define a discrete loss function f by assigning a nonnegative numerical value to each cell. The discrete loss function can be shown as a matrix, as in the following example:

		$\widehat{y}_t$			
		L-	S-	S+	L+
y_t	L-	0	1	2	3
	S-	1	0	2	3
	S+	3	2	0	1
	L+	3	2	1	0

(1)

with L-, S-, S+, L+ being the $(-\infty, -l)$, $(-l, 0)$, $(0, +l)$, $(+l, +\infty)$ intervals, respectively, for a given constant l , conveniently chosen by the user.[4] Le us denote by a the k-vector $(k \leq m^2)$ of possible losses associated to the different cells, i.e., the k-vector of possible values of f, ordered increasingly. In the example , $a = (0, 1, 2, 3)$. We are associating a high loss to incorrectly forecasting the sign, while the magnitude of the forecast error is of secondary importance. A forecast with the right sign always receives in (1) a penalty lower than any forecast with the wrong sign, with independence of the size of the forecast error. This is a loss structure that may be reasonable in many applications, although many other alternative choices for f would also be admissible.

A discrete loss function presents significant advantages for the evaluation of macroeconomic and financial forecasts: i) by appropriately choosing the number of elements in the partition and the associated penalties, the user can accommodate the choice of f to each

[4] Here we use the same partition for the data and the forecasts, although the analysis can be extended without any difficulty to the case when the two partitions are different.

specific forecasting context, *ii*) at a difference from most standard loss functions, which are usually a function of just the forecast error, a discrete loss allows for a rich forecast evaluation that it can pay attention to a variety of characteristics of data and forecasts, *iii*) a discrete loss function can take into account the signs as well as the size of both, data and forecast, which allows for a simple incorporation of different types of asymmetries,[5] *iv*) a discrete loss imposes an upper bound on the loss function, thereby reducing or even eliminating the distorting effect of outliers when evaluating the performance of a set of forecasts, *v*) a discrete loss function is a natural choice for the evaluation of forecasts of qualitative variables.

Besides the notional characteristics we described above in favor of discrete loss functions as an interesting choice for forecast evaluation, they also possess two very significant technical advantages, as we are about to see.

3.2 The DISC test

We propose a simple statistical test of the null hypothesis that forecasts are not useful under a discrete loss function f , by testing $H_0 \equiv E(f_t) = E(f_t^{IE})$ against $H_1 \equiv E(f_t) < E(f_t^{IE})$. The appropriate test statistic is the difference between the sample means $\overline{f}$ and $\overline{f}^{IE}$, as unbiased estimates of $E(f_t)$ and $E(f_t^{IE})$, and the test will be based on the asymptotic distribution of the $\overline{f} - \overline{f}^{IE}$ statistic. Under most loss functions, the practical implementation of the test would face two difficulties. First, we lack the sample observations on f_t^{IE} needed to compute $\overline{f}^{IE}$. If the specification of the loss function allows for an adequate characterization of $E(f_t^{IE})$ as a function of data and forecasts, the unbiased estimator $\overline{f}^{IE}$ will be obtained substituting sample estimates in that expression. For instance, for the square forecast error loss $f(y_t, \widehat{y}_t) = (y_t - \widehat{y}_t)^2$ we would have: $E(f_t^{IE}) = E[(y_t - \widehat{y}_t)^2] = E(y_t^2) + E(\widehat{y}_t^2) - 2E(y_t)E(\widehat{y}_t)$, and substituting sample averages for the mathematical expectations, we could compute the value of $\overline{f}^{IE}$. While this argument will not work for every continuous loss function f , the mean value $\overline{f}^{IE}$ can always be obtained under a discrete loss f .[6] Second, even if we knew the analytical expression for $E(f_t^{IE})$, the asymptotic distribution of $\overline{f} - \overline{f}^{IE}$ would generally not be easy or even feasible to obtain for continuous loss functions.[7] Luckily enough, this is again not a problem when working with a discrete loss f , as we now show.

Under a discrete loss f , we have $E(f_t) = ap$ and $E(f_t^{IE}) = ap^{IE}$, with $p(r) = P(f_t = a(r))$ and $p^{IE}(r) = P\left[f_t = a(r) \mid (y_t, \widehat{y}_t) \textit{ being stochastically independent}\right]$. On the other hand, $p(r) = \sum_{(i,j)\in C(r)} p_{ij}$ and $p^{IE}(r) = \sum_{(i,j)\in C(r)} p_{ij}^{IE}$, with p_{ij}^{IE} being the probability that data and forecasts fall in the (i,j)-cell if they are stochastically independent, while $C(r)$ represents the set of all quadrants where f takes the value $a(r)$. By definition of stochastic independence, we have $p_{ij}^{IE} = p_{i.}p_{.j}$ and we can easily get the expression for $E(f_t^{IE})$. Its estimator $\overline{f}^{IE}$ is defined substituting in that expression the estimator $\widehat{p}_{ij}^{IE} = \widehat{p}_{i.}\widehat{p}_{.j}$ for p_{ij}^{IE}. It is the discrete nature of the loss function f that allows us to define an estimator $E(f_t^{IE})$ that can be easily calculated from the sample relative frequencies.

We can now proceed to describing our test proposal. Let $P = (p_{11}, p_{12}, ..., p_{1m}, p_{21}, ..., p_{2m}, ..., p_{m1}, ..., p_{mm})'$ be the m^2-column vector that contains the theoretical probabilities p_{ij} for the quadrants associated to the partition of the space of data and forecasts, and $\widehat{P}$ its maximum likelihood esti-

[5] Which are so natural in Economics. For instance, it is hard to believe that an investor will regret getting a return much higher than it was predicted when a financial asset was bought.

[6] To characterize the expression for $E[f(y_t, \widehat{y}_t)]$ under independence of y_t and $\widehat{y}_t$, we need the function f to be of the form: $f(y_t, \widehat{y}_t) = \sum_k a_k h_k(y_t) g_k(\widehat{y}_t)$ for any set of constants a_k and functions h_k, g_k.

[7] For instance, if $f = (y_t - \widehat{y}_t)^2$, we would need to characterize the asymptotic distribution of the sample statistic, $2T^{-1}(T^{-1}\sum y_t \sum \widehat{y}_t - \sum y_t \widehat{y}_t)$.

mator, based on relative frequencies. Using standard results, we have $\sqrt{T}(\widehat{P}-P) \xrightarrow{L} N(0, V_P)$, with $V_P = \Omega - PP'$ and Ω a diagonal $m^2 \times m^2$ matrix with the elements of P along the diagonal. Let us know consider the differentiable function $\varphi(.)$ of R^{m^2} on R^k defined by: $\varphi(P) = p - p^{IE} = \left(\sum_{C(1)} p_{ij} - p_{i.}p_{.j}, ..., \sum_{C(k)} p_{ij} - p_{i.}p_{.j} \right)'$. Putting together both results we have: $\sqrt{T}\left[(\widehat{p} - \widehat{p}^{IE}) - (p - p^{IE})\right] \xrightarrow{L} N(0, \nabla\varphi(P) V_P \nabla\varphi(P)')$, with $\nabla\varphi(P)$ being the $k \times m^2$ Jacobian matrix for the vector function $\varphi(P)$. Finally, multiplying by a we have the asymptotic distribution of $\overline{f} - \overline{f}^{IE}$ under the null hypothesis $E(f_t) = E(f_t^{IE})$: $\sqrt{T}(\overline{f} - \overline{f}^{IE}) = a(\widehat{p} - \widehat{p}^{IE}) \xrightarrow{L} N(0, G_p)$, with $G_p = a\nabla\varphi(P) V_P \nabla\varphi(P)' a'$.

We use the consistent estimator $\widehat{G}_p$, which allows us to maintain the same limiting distribution. Therefore, the proposed DISC test for the null H_0 against H_1 is:

$$D = \sqrt{T}\widehat{G}_p^{-1/2} a(\widehat{p} - \widehat{p}^{IE}) \underset{H_0}{\xrightarrow{L}} N(0,1), \tag{2}$$

with $\widehat{G}_p = a\left[\nabla\varphi(P)\right]_{P=\widehat{P}} \widehat{V}_P \left[\nabla\varphi(P)\right]'_{P=\widehat{P}} a'$ and $\widehat{V}_P = V_P|_{P=\widehat{P}}$. The expression for matrix $\nabla\varphi(P)$ is given in Appendix A. The critical region for the test corresponds to the lower tail of the $N(0,1)$ distribution.

Obviously, the test will be invariant to application of a scale factor λ on the loss function. It is not difficult to show that the one sided version of the PT test, where the critical region is just the upper-tail of the distribution, is a special case of the DISC test when there are only two values in a, one of them being the penalty assigned to every cell along the main diagonal in the loss matrix, and another one for the rest of the cells, the first value being smaller than the second one.

4 Applying the DISC test

4.1 Back to Example 1

Let us get back to Example 1 in Section 2 to illustrate the behavior of the DISC test. We start by implementing the test for the three forecasting results M1, M2 and M3, under two alternative discrete loss functions, the one defined by (1), and an alternative one characterized by:

		$\widehat{y}_t$			
		G-	P-	P+	G+
y_t	G-	0	1.75	2	3
	P-	1.75	0	2	3
	P+	3	2	0	1.75
	G+	3	2	1.75	0

(3)

The difference between (1) and (3) reduces to the loss associated to forecasts that have the right sign but the wrong magnitude, which is equal to 1 in (1) while being 1.75 in (3). Therefore, while (1) assigns high value, i.e., a low loss, to predicting the sign correctly, the difference under (3) between the losses associated to a correct and an incorrect prediction of the sign is small, provided the size of the error is not too large.

Figure 1. Test results for models M1, M2 and M3

Losses	Function **(1)**			Function **(3)**		
	Observed value for the test statistic					
	CT	PT	DISC	CT	PT	DISC
M1	292.31	-353.55	40.62	292.31	-353.55	33.46
M2	292.31	-353.55	-17.60	292.31	-353.55	2.56
M3	292.31	74.94	-43.77	292.31	74.94	-48.95
	p-value[8]					
	CT	PT	DISC	CT	PT	DISC
M1	0.0	1.0	1.0	0.0	1.0	1.0
M2	0.0	1.0	**0.0**	0.0	1.0	**0.99**
M3	0.0	0.0	0.0	0.0	0.0	0.0

Figure 1 displays the results of applying the DISC test as well as the CT and PT tests to the three hypothetical forecasting models. In the three tests, the null hypothesis is that the set of forecasts is not useful. At a difference from the CT and PT tests, the DISC test associates different values to the test statistic for forecasts M1 and M2. For model M3, the three tests correctly reject the null hypothesis. For M1, CT also rejects the null, incorrectly, while the PT and DISC tests do not reject it. Finally, and this is the most relevant case, the CT and PT tests lead to a given decision on the lack of utility of M2 forecasts with independence of the forecasting context in which the observed frequencies were obtained, which is not too reasonable as already discussed in Section 2. On the other hand, the DISC test rejects the null hypothesis in favor of considering that the predictions are useful if the loss function is (1), while considering the set of forecasts not to be useful if the loss function is given by (3). We consider that to be a reasonable behavior for a forecast evaluation test in a situation like that summarized by M2: to reject the hypothesis of lack of utility of the forecasts if and only if correctly forecasting the sign is relevant enough.

As we see in this application, DISC solves some of the limitations of the CT and PT tests pointed out in Section 2 and in the Introduction. It does not make unacceptable mistakes as rejecting the null hypothesis of lack of utility of forecasts when they present a negative correlation with the data. Even much more importantly, the test decision will depend on the definition of 'useful forecasts' made by the user in each specific application of the test through the choice of loss function, and the DISC test can accommodate any level of desired detail in that definition.

To analyze this last statement in depth, we now illustrate the sensitivity of the DISC test to the numerical values chosen for the loss function. To that end, we perform five simple exercises in which we apply the DISC test under the matrix of joint frequencies M2 with different loss functions. The DISC test would have rejected the null hypothesis under M3 for any reasonable loss function, while not rejecting the null hypothesis under M1. On the other hand, the result of the test under M2 depends on the numerical values of the loss function, as it is to be desired. Remember that what is distinctive about M2 forecasts is that they always have the right sign, but they are never highly precise. To consider a wide array of possibilities, we use the pattern defined by matrix (4) under the constraint $\lambda_3 > \lambda_2 > \lambda_1 > 0$. This way, we penalize more heavily a forecast that has the wrong sign than one with the right sign, independently of the size of the error in each case, but we also take into account the size of the forecast error. Furthermore, we guarantee some symmetry in the loss function,

[8] The critical region for tests CT and DISC is the upper and lower tail, respectively, of their corresponding test distributions. The critical region for the PT includes both tails.

with the same numerical loss for quadrants like (G-, P+) and (G+, P-). This pattern has already been used in (1) and (3).

$$
\begin{array}{cc|cccc}
 & & \multicolumn{4}{c}{\widehat{y}_t} \\
 & & \text{G-} & \text{P-} & \text{P+} & \text{G+} \\
\hline
 & \text{G-} & 0 & \lambda_1 & \lambda_2 & \lambda_3 \\
y_t & \text{P-} & \lambda_1 & 0 & \lambda_2 & \lambda_3 \\
 & \text{P+} & \lambda_3 & \lambda_2 & 0 & \lambda_1 \\
 & \text{G+} & \lambda_3 & \lambda_2 & \lambda_1 & 0
\end{array}
\tag{4}
$$

a) As a first exercise, we take $\lambda_1 = 1$, $\lambda_2 = 2$, and apply the DISC test for values of λ_3 in the interval $(2, 3)$.The second exercise is similar, taking $\lambda_1 = 1$, $\lambda_3 = 3$ while λ_2 takes values in the interval $(1, 3)$. In both exercises, the p-value associated to the DISC test is 0.0 in all possible cases, so that the test would always reject the null hypothesis that the forecasts from M2 are not useful.

b) We might be tempted to conclude from the previous exercise that the DISC test will always reject the null hypothesis that the forecasts from M2 are not useful under any loss matrix (4) verifying $\lambda_3 > \lambda_2 > \lambda_1 = 1$, but that is not the case. Even under that constraint, if λ_2 and λ_3 are both close enough to λ_1, then the DISC test might not reject the null hypothesis, because the penalty associated to forecasts with the right sign but the wrong size (λ_1) is not too different from the loss associated to any forecast that has the wrong sign (λ_2 or λ_3). If we maintain $\lambda_1 = 1$ and let λ_2 and λ_3 vary inside the intervals $(1, 2)$ and $(2, 3)$, respectively, the p-value is above 0.05 in some cases, like when $\lambda_2 = 1.05$ and $\lambda_3 < 2.15$, or when $\lambda_2 = 1.10$ and $\lambda_3 < 2.05$.

The two previous points show that a clear distinction between the losses associated to forecasts with the wrong sign and some of the forecasts with the right sign is needed for the DISC test to conclude in favor of the usefulness of the M2 forecasts, which seems a desirable condition.

c) The most interesting situations are those in which λ_1 is let to change, since this parameter defines the only loss made by model M2 and hence, it is the most decisive to understand the behavior of the test. So, we now maintain $\lambda_2 = 2$ and $\lambda_3 = 3$, while λ_1 takes values in the interval $(0, 2)$, with the results shown in Figure 2a. The p-value is 0.0 for values of λ_1 up to $\lambda_1 = 1.59$, rapidly increasing to reach 1.0 for $\lambda_1 = 1.76$. This is consistent with the result obtained using (3) as loss function, and emphasizes again that the M2 forecasts will be seen as useful only if there is a substantial difference in value between forecasts with the wrong and the right sign.

d) Finally, to complete the analysis in *c*), we perform another exercise letting λ_1 and λ_2 vary inside the intervals $(0, 2)$ and $(2, 3)$, respectively, while $\lambda_3 = 3$. The DISC test rejects the null hypothesis whenever $\lambda_1 < 1.6$, in coherence with the results obtained in *c*). If $\lambda_1 > 1.6$, then the decision of the DISC test for M2 will depend on the value of λ_2. Once again, the closer are λ_1 and λ_2 to each other, the less relevant will be forecasting the right sign and hence, the more likely will be not to reject the null hypothesis. In Figure 2b we show the p-values for some values of λ_1 and λ_2. Specifically, we draw the curves of p-values as λ_2 changes for some fixed values of λ_1, all of them above 1.6. On the other hand, in Figure 3 we present the combinations (λ_1, λ_2) for which we obtained a p-value equal to 0.01, which allows us to gain some intuition as to the level of the λ_1/λ_2 ratio below which the DISC test will reject the null hypothesis of lack of usefulness of the M2 forecasts under a loss matrix (4) with $\lambda_3 = 3$.

Figure 2. p-values of the DISC test for example M2, under a loss (4)

Figure 2.a. $\lambda_2 = 2$, $\lambda_3 = 3$

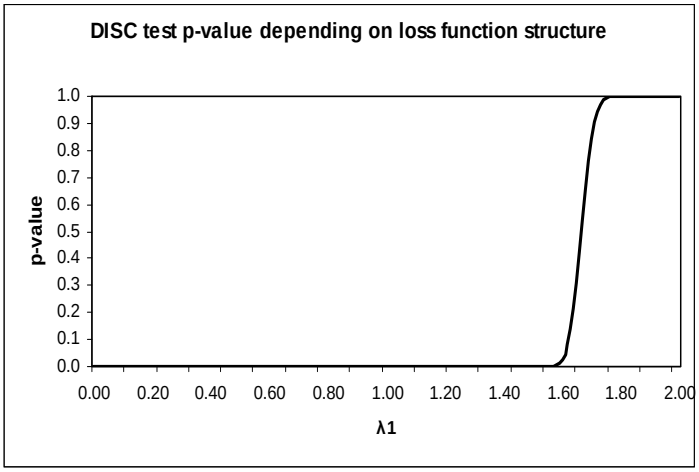

Figure 2.b. $\lambda_3 = 3$

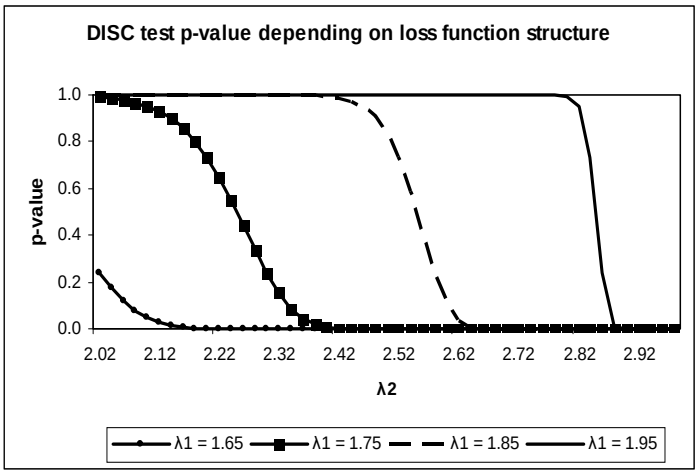

Figure 3. (λ_1, λ_2) combinations such that DISC p-value = 0.01 for M2 forecasts under (4), with $\lambda_3 = 3$.

	λ_1			
	1.65	1.75	1.85	1.95
λ_2	2.14	2.40	2.64	2.88
λ_1/λ_2	0.77	0.73	0.70	0.68

For each value of λ_1, the p-value of the DISC test for M2 will be below 0.01 whenever λ_2 is higher than the value associated to λ_1 .

4.2 Simulation results

4.2.1 Experimental design

To obtain further evidence on the different behavior of the DISC test and the CT and PT tests, we now perform a simulation experiment. We will sample T $(y_t, \widehat{y}_t)$ pairs from a Bivariate Normal distribution with zero means, correlation coefficient ρ and unit variances, and apply the three tests. The first variable will be considered as the data and the second variable as the forecasts. We will use values: $\rho = 0,\ 0.4,\ 0.75,\ 0.9$ and $T = 10,\ 25,\ 50$.[9] The numerical value of ρ will allow us to control if the set of forecasts are useful. The test will employ a 4×4 partition of the R^2-space, based on intervals $(-\infty, -l)$, $(-l, 0)$, $(0, +l)$, $(+l, +\infty)$, with $l = 0.8416$. Furthermore, the DISC test will use (1) as loss function. Under $l = 0.8416$, the marginal probabilities that the data fall in each of the intervals are 0.20, 0.30, 0.30 and 0.20, respectively, and the same applies to the forecasts, which looks reasonable. We repeated the simulation exercises with $l = 0.5244$, those probabilities then becoming 0.30, 0.20, 0.20 and 0.30, to obtain the same qualitative results.

This analysis is sort of the opposite to the one carried out at the end of the previous section. There, we kept the matrix of frequencies M2 fixed, i.e., there was only one set of forecasts set, while we were changing the definition of the discrete loss function. By contrast, we now vary the set of forecasts while maintaining always the same discrete loss function.

Before proceeding, it is crucial to understand that in our framework we cannot perform a standard analysis of size and power. The null hypothesis for the tests is that the set of forecasts lacks usefulness. Therefore, even though each test defines that hypothesis in a specific manner, the null hypothesis is ambiguous in nature, and it will usually not be possible to know before hand whether it is true or false, in spite of the fact that we are running a simulation experiment.

We proceed as follows:

a) if $\rho = 0$, it is clear that the set of forecasts is not useful and the null hypothesis should not be rejected. This is a standard size exercise.

b) if ρ is high enough, the forecasts should be considered useful and the null hypothesis should be rejected, and we can analyze the power of the tests in a standard sense. In our experiment, this will apply to the case $\rho = 0.9$.

c) if ρ takes an intermediate value, as it might be expected in practical applications, we cannot conclusively say whether the forecasts are useful. This will be the case in our experimental design when $\rho = 0.4$. We will then study each sample realization and check whether the decision taken by each test looks 'reasonable' given the partition and loss function that have been defined.

The case when $\rho = 0.75$ can be interpreted as either *b*) or *c*), so we will apply both types of analysis to that case.

4.2.2 Results

Table 1 presents the rejection probabilities for the three tests. We can compare the performance in size of the three tests (when $\rho = 0$) and their performance in terms of power (when $\rho = 0.9$). All tests are reasonably unbiased in size (except for the CT test when $T = 10$), DISC being the test with the highest power for small sample sizes. If we take the view that $\rho = 0.75$ must be interpreted as an exercise in power, i.e., that forecasts that have correlation of $\rho = 0.75$ with the data should be seen as useful, then the results in Table 1 are even more evident in favor of the DISC test being more powerful than the CT and PT alternatives.

[9] We restrict our attention to samples of length $T \leq 50$, since the case $T > 50$ does not usually arise in practice.

Table 1. Rejection probabilities (%). $\alpha = 5\%$.

$(y_t, \widehat{y}_t) \sim N(0_{2\times 1}, \Sigma)$, $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$

ρ	T	CT 4×4	PT	DISC
0.00	10	1.7	6.1	8.4
	25	4.1	4.5	6.5
	50	4.4	4.2	5.3
0.40	10	3.1	17.6	29.4
	25	13.5	25.3	45.7
	50	33.3	40.0	69.2
0.75	10	11.5	48.2	68.2
	25	71.2	78.5	95.0
	50	99.0	96.8	99.8
0.90	10	32.8	80.0	89.7
	25	97.9	98.8	99.9
	50	100.0	100.0	100.0

Number of realizations: 5000.

In situations with an intermediate degree of correlation between data and forecasts, the analysis of size and power does not apply. For $\rho = 0.4$, Table 1 shows again that the DISC test rejects the null hypothesis of lack of utility of the set of forecasts more often than CT and PT. But, since we do not know a priori whether or not the set of forecasts is then useful, such a result is hard to interpret. Because of that, we have analyzed each of the 5000 samples produced under this design, with the intention of checking how reasonable were the decisions made by each test. We have paid attention to those simulations in which the CT and PT tests take the same decision, while the DISC test takes the opposite decision. The percentage of simulations when this circumstance arises for the $\rho = 0.4$ and $\rho = 0.75$ designs is given in Table 2. We can see that it is very unlikely that the DISC test will not reject the null hypothesis of lack of utility of the set of forecasts whenever the CT and PT tests reject it. So, the discrepancy between the tests arises in the opposite situation. Under the loss function (1), it is relatively frequent that the DISC test rejects the null hypothesis at the same time that the CT and PT tests do not reject it. We will call these 'type-R simulations'. As shown in Table 2, such a probability falls between 10% and 23%, except in the case $\rho = 0.75$, $T = 50$, when the three tests reject the null hypothesis in almost 100% of the samples.

Table 2. Estimated probability (%)
that the DISC test will take a decision
contrary to the CT and PT tests

ρ	T	CT and PT: NR DISC: R	CT and PT: R DISC: NR
0.40	10	14.6	0.2
	25	20.3	0.4
	50	22.8	0.4
0.75	10	21.3	0.2
	25	10.1	0.1
	50	0.3	0.0

R and NR denote rejection and not rejection of H_0, respectively.

Table 3 summarizes the results from type-R simulations. For each (ρ, T)-pair we select three specific type-R simulations: those corresponding to the maximum, minimum and median value of $1 - v$, where v refers to the p-value for the DISC test. We will denote those simulations by *smax, smin* and *smedian*, respectively. We could interpret this choice as selecting the cases when the discrepancy between DISC and the other two tests was largest (maximum $1 - v$), lowest (minimum $1 - v$) and an intermediate case (median $1 - v$).[10] For each of these three simulations, we present in Table 3 the sample relative frequencies for the four possible loss values according to (1).

The conclusion that the DISC test made the right decision is less controversial for small samples. For instance, if $T = 10$, we have simulations like *smax*, where the forecasts have always been correct in sign. Furthermore, when $T = 10$, forecasts have also been correct in size at least 40% of the times for both $\rho = 0.40$ as well as for $\rho = 0.75$. Yet, CT and PT will not reject the lack of utility of the forecasts, while DISC does reject it. Using *smedian* and *smin* simulations the situation is less extreme, but with 80% of forecasts having the right sign, it should be easy to argue that the DISC test still leads to the right decision in these simulations. The situation gradually becomes less clear as T increases since CT and PT work better then. But if we revise each one of the simulations in Table 3, it is hard to detect a case in which the DISC test makes a decision that we could consider unreasonable. The more arguable case might be the *smin* simulation for $\rho = 0.4$, $T = 50$. There, the forecasts had the right sign 56% of the times, but only 36% of the times were also right in size, while among the 44% of forecasts that missed the sign, only 12% forecasts had the wrong sign and the wrong size. But, even in this case, the decision made by DISC to reject that the forecasts are not useful looks acceptable, according to the loss function (1).

[10] Other criterions lead to similar conclusions. That would be the case if we used the difference between the mean of the p-values obtained for the CT and PT test and v, or if we used the numerical value of $\overline{f}^{IE} - \overline{f}$.

Table 3. Detailed information on representative type-R simulations

ρ	T		$\overline{f}$	$\overline{f}^{IE}$	$\widehat{p}(0)$	$\widehat{p}(1)$	$\widehat{p}(2)$	$\widehat{p}(3)$
0.40	10	*smax*	0.60	1.53	0.40	0.60	0.00	0.00
		smedian	0.80	1.47	0.50	0.30	0.10	0.10
		smin	0.70	1.28	0.50	0.30	0.20	0.00
	25	*smax*	0.84	1.48	0.40	0.40	0.16	0.04
		smedian	0.96	1.36	0.44	0.28	0.16	0.12
		smin	1.04	1.35	0.36	0.32	0.24	0.08
	50	*smax*	0.96	1.41	0.34	0.40	0.22	0.04
		smedian	1.10	1.41	0.36	0.28	0.26	0.10
		smin	1.20	1.46	0.36	0.20	0.32	0.12
0.75	10	*smax*	0.50	1.34	0.50	0.50	0.00	0.00
		smedian	0.70	1.30	0.50	0.30	0.20	0.00
		smin	1.00	1.56	0.30	0.50	0.10	0.10
	25	*smax*	0.84	1.50	0.36	0.44	0.20	0.00
		smedian	0.96	1.43	0.40	0.32	0.20	0.08
		smin	1.00	1.32	0.32	0.36	0.32	0.00
	50	*smax*	0.92	1.37	0.40	0.30	0.28	0.02
		smedian	0.96	1.30	0.42	0.24	0.30	0.04
		smin	1.08	1.35	0.36	0.30	0.24	0.10

$\widehat{p}(i)$: sample relative frequency for the event $f_t = i$.

To summarize the analysis in this section, we can say that the DISC test is more powerful than the CT and PT tests in those situations in which the set of forecasts is clearly useful (high values of ρ). In cases when it is unclear a priori whether or not the set of forecasts is useful (intermediate values of ρ), we can at least say that the DISC test always behaves reasonably, according to the definition of the chosen loss function. In such situation, we must see the decisions reached by the CT and PT tests as arbitrary, since it would be unclear, by looking just at the relative frequencies of forecasts with the right or wrong sign and size, that the set of forecasts is useful. By contrast, by paying attention at the information provided by each data point and the associated forecast, the DISC test makes better decisions.

5 Conclusions

We have analyzed three non parametric tests to evaluate the quality of a set of point forecasts, which can be used even if we ignore the probability distributions of data and forecasts. Two of them are standard in the literature, the Contingency Table test (CT) and the Pesaran and Timmermann test (1992) (PT), while we have introduced the DISC test. We have shown how the CT test can easily make unacceptable mistakes even in situations where the forecasts are obviously not useful. Furthermore, given a set of numerical forecasts and data, the conclusion of the CT and PT tests is independent of the particular application in which the data and forecasts have been generated, with a suboptimal performance in many forecasting contexts.

The problem arises because both tests focus just on the independence or lack thereof between data and forecasts, an approach which precludes a finer evaluation of each $(data, forecast)$-

pair and essentially leads to a rigid definition of what we understand by useful forecasts. They are also based on the joint sample frequency distribution of data and forecasts, without fully exploiting the information in their numerical values. On the contrary, the DISC test is based on a discrete loss function that characterizes what is meant by useful forecasts in each specific application, solving the mentioned limitations of alternative tests like the CT and PT tests. Discrete loss functions are interesting in many practical situations, as it is the case when a correctly signed forecast is particularly important or when forecasting qualitative data. Discrete loss functions are very flexible, since they do not need to have the forecast error as their only argument. That way, it is very easy to accommodate any type of asymmetry in the valuation of forecast errors, which permits a richer evaluation of forecasts. Besides, the discrete nature of the function allows us to obtain the probability distribution for the DISC test statistic, which could not be found under general continuous loss functions.

Our results suggest that the DISC test performs better than the two standard tests: it does not make unacceptable mistakes like those occasionally made by CT, and it seems to be more powerful in situations when the set of forecasts is clearly useful. In experimental designs when there is ambiguity about the utility of the set of forecasts, the behavior of the DISC test is at least reasonable, according to the utility criterion that the user may have established through the numerical specification of the discrete loss function.

A Appendix: The expression for $\nabla\varphi(P)$

Remember that the function $\varphi(P)$ is

$\varphi(P) = \left(\sum_{C(1)} p_{uv} - p_{u.}p_{.v}, ..., \sum_{C(k)} p_{uv} - p_{u.}p_{.v} \right)'$ and that $C(r)$ is the set of (u,v)-quadrants where f takes the value a_r. We want to obtain the expression for the $k \times m^2$ matrix $\nabla\varphi(P) = \begin{pmatrix} \frac{\partial\varphi_1}{\partial p_{11}} & \frac{\partial\varphi_1}{\partial p_{12}} & ... & \frac{\partial\varphi_1}{\partial p_{mm}} \\ \frac{\partial\varphi_2}{\partial p_{11}} & \frac{\partial\varphi_2}{\partial p_{12}} & ... & \frac{\partial\varphi_2}{\partial p_{mm}} \\ ... & ... & ... & ... \\ \frac{\partial\varphi_k}{\partial p_{11}} & \frac{\partial\varphi_k}{\partial p_{12}} & ... & \frac{\partial\varphi_k}{\partial p_{mm}} \end{pmatrix}$, where φ_r is the r-th element of $\varphi(P)$, i.e., $\varphi_r = \sum_{C(r)} p_{uv} - p_{u.}p_{.v}$.

Before giving the general expression for $\frac{\partial\varphi_r}{\partial p_{ij}}$, let us work with a particular example which may help the reader to understand the ongoing general expression easier:

Consider the 4×4 loss function:

		$\widehat{y}_t$			
		r_1	r_2	r_3	r_4
y_t	r_1	0	1	2	3
	r_2	1	0	2	3
	r_3	3	2	0	1
	r_4	3	2	1	0

and let us see how to calculate the derivative $\frac{\partial\varphi_2}{\partial p_{31}}$. The function is $\varphi_2 = \sum_{C(2)} p_{uv} - p_{u.}p_{.v}$. The set $C(2)$ consists of those quadrants with a loss $a_2 = 1$, i.e., $C(2) = \{(1,2),(2,1),(3,4),(4,3)\}$. Therefore φ_2 takes the expression $\varphi_2 = (p_{12} - p_{1.}p_{.2}) + (p_{21} - p_{2.}p_{.1}) + (p_{34} - p_{3.}p_{.4}) + (p_{43} - p_{4.}p_{.3})$.

We should find those terms that include the parameter p_{31}. As the marginal probabilities $p_{i.}$ and $p_{.j}$ are the sum of the i-th row and j-th column probabilities, respectively, the parameter p_{31} implicitly appears in $p_{3.}$ and $p_{.1}$. As p_{31} appears in $p_{3.}$ and $p_{.1}$, the derivative $\frac{\partial\varphi_2}{\partial p_{31}}$ is $\frac{\partial\varphi_2}{\partial p_{31}} = d_{31}^{(2)} = -p_{2.} - p_{.4}$. Had the $(3,1)$-quadrant also been included in the set $C(2)$,

p_{31} would have been the first element of another term into brackets, and the derivative would have been $1 - d_{31}^{(2)}$.

We are now prepared to understand the general expression for $\frac{\partial \varphi_r}{\partial p_{ij}}$:

$\frac{\partial \varphi_r}{\partial p_{ij}} = d_{ij}^{(r)} = -\left(\sum_{(i,v) \in C(r)} p_{\cdot v} + \sum_{(u,j) \in C(r)} p_{u \cdot} \right)$, if the (i,j)-quadrant is not included in $C(r)$, and $\frac{\partial \varphi_r}{\partial p_{ij}} = 1 - d_{ij}^{(r)}$, otherwise.

References

[1] Ash, J.C.K., Smyth, D.J and Heravi, S.M. (1998). Are OECD Forecasts Rational and Useful?: a Directional Analysis, *International Journal of Forecasting* 14, 381-391.

[2] Greer, M. (2003). Directional Accuracy Tests of Long-Term Interest Rate Forecasts, *International Journal of Forecasting* 19, 291-298.

[3] Henriksson, R.D., Merton and R.C. (1981). On Market Timing and Investment Performance. II: statistical procedures for evaluating forecasting skills, *Journal of Business* 54, 513-533.

[4] Joutz, F. and Stekler, H.O. (2000). An Evaluation of the Predictions of the Federal Reserve, *International Journal of Forecasting* 16, 17-38.

[5] Kolb, R.A. and Stekler, H.O. (1996). How well do Analysts Forecast Interest Rates?, *Journal of Forecasting* 15, 385-394.

[6] Leitch, G. and Tanner, J.E. (1995). Professional Economic Forecasts: Are they Worth their Costs?, *Journal of Forecasting* 14, 143-157.

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

[8] Mills, T.C. and Pepper, G.T. (1999). Assessing the Forecasters: an Analysis of the Forecasting Records of the Treasury, the London Business School and the National Institute, *International Journal of Forecasting* 15, 247-257.

[9] Oller, L. and Bharat, B. (2000). The Accuracy of European Growth and Inflation Forecasts, *International Journal of Forecasting* 16, 293-315.

[10] Pesaran, M.H. and Timmermann, A. (1992). A Simple Nonparametric Test of Predictive Performance, *Journal of Business and Economic Statistics* 10, 461-465.

[11] Pons, J. (2000). The Accuracy of IMF and OECD Forecasts for G7 Countries, *Journal of Forecasting* 19, 53-63.

[12] Schnader, M.H. and Stekler, H.O. (1990). Evaluating Predictions of Change, *The Journal of Business* 63, 1, 99-107.

[13] Stekler, H.O. (1994). Are Economic Forecasts Valuable?, *Journal of Forecasting* 13, 495-505.