On Forecast Evaluation

Por: Wilmer Osvaldo Martínez-Rivera,
Manuel Dario Hernández-Bejarano,
Juan Manuel Julio-Román

Borradores de
ECONOMÍA

Núm. 825
2014

tá - Colombia - BANCO DE LA REPÚBLICA COLOMBIA - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Col

On Forecast Evaluation[*]

Wilmer Osvaldo Martínez-Rivera[†]
Manuel Dario Hernández-Bejarano[‡]
Juan Manuel Julio-Román[§]

Abstract

We propose to assess the performance of k forecast procedures by exploring the distributions of forecast errors and error losses. We argue that non systematic forecast errors minimize when their distributions are symmetric and unimodal, and that forecast accuracy should be assessed through stochastic loss order rather than expected loss order, which is the way it is customarily performed in previous work. Moreover, since forecast performance evaluation can be understood as a one way analysis of variance, we propose to explore loss distributions under two circumstances; when a strict (but unknown) joint stochastic order exists among the losses of all forecast alternatives, and when such order happens among subsets of alternative procedures. In spite of the fact that loss stochastic order is stronger than loss moment order, our proposals are at least as powerful as competing tests, and are robust to the correlation, autocorrelation and heteroskedasticity settings they consider. In addition, since our proposals do not require samples of the same size, their scope is also wider, and provided that they test the whole loss distribution instead of just loss moments, they can also be used to study forecast distributions as well. We illustrate the usefulness of our proposals by evaluating a set of real world forecasts.

Keywords: Forecast evaluation, Stochastic order, Multiple comparison.
JEL: C53, C12, C14.

[*] **First Draft for comments, please do not circulate**. We are grateful to Daniel Moreno-Cortés for his careful proofreading. However, any remaining errors as well as the conclusions and opinions contained in this paper are the sole responsibility of its authors and do not compromise BANCO DE LA REPUBLICA, its Board of Governors or Universidad Nacional de Colombia.

[†] wmartiri@banrep.gov.co, Specialized Professional, Statistics Division, Economic Research VP, BANCO DE LA REPUBLICA, Bogotá, Colombia.

[‡] mhernabe@banrep.gov.co, Specialized Professional, Statistics Division, Economic Research VP, BANCO DE LA REPUBLICA, Bogotá, Colombia.

[§] *Corresponding author:* jjulioro@banrep.gov.co, Senior Researcher, Research Unit, Technical Presidency, BANCO DE LA REPUBLICA, and part time Associate Professor, Department of Statistics, Universidad Nacional de Colombia, Bogotá, Colombia.

Sobre la Evaluación de Pronósticos[§]

Wilmer Osvaldo Martínez-Rivera[1]
Manuel Dario Hernández-Bejarano[2]
Juan Manuel Julio-Román[3]

Resumen

Proponemos evaluar el desempeño de k procedimientos de pronóstico explorando las distribuciones de los errores de pronóstico y de sus pérdidas. Argumentamos que los errores no sistemáticos de pronóstico se minimizan cuando su distribución es simétrica y unimodal, y que la precisión de los pronósticos debe evaluarse a través del orden estocástico de sus pérdidas en vez del orden de las pérdidas esperadas, que es como se propone en trabajos anteriores. Adicionalmente, como la evaluación de pronósticos se puede entender como un análisis de varianza a una vía, proponemos explorar las distribuciones de las pérdidas bajo dos circunstancias; cuando hay un orden estocástico conjunto (desconocido) entre las pérdidas de los k procedimientos, y cuando este orden ocurre en subconjutos de estos. A pesar de que el orden estocástico es más fuerte que el orden de las pérdidas esperadas, nuestras propuestas son tan potentes como las competidoras además de ser robustas a las correlaciones, autocorrelaciones y heterogeneidades consideradas para estas. De igual manera, como nuestras propuestas no requieren muestras del mismo tamaño, su campo de aplicación es más amplio, y como exploran la distribución de la pérdida en vez de solo sus momentos, también se pueden utilizar para evaluar las distribuciones de pronóstico de distintos procedimientos. Finalmente, ilustramos la utilidad de nuestras propuestas evaluando un conjunto de pronósticos de la vida real.

Palabras Clave: Evaluación de pronósticos, Orden estocástico, Comparaciones múltiples.
JEL: C53, C12, C14.

[§] **Primera Versión para comentarios, por favor no circular**. Los autores agradecen a Daniel Moreno Cortéz por su cuidadosa prueba de lectura. Sin embargo, cualquier error que persista en este escrito, asi como sus conclusiones y recomendaciones son responsabilidad exclusiva de sus autores y no comprometen al BANCO DE LA REPUBLICA, su Junta Directiva o la Universidad Nacional de Colombia.

[1] wmartiri@banrep.gov.co, Profesional Especializado, División de Estadística, Sub Gerencia de Estudios Económicos, BANCO DE LA REPUBLICA, Bogotá, Colombia.

[2] mhernabe@banrep.gov.co, Profesional Especializado, División de Estadística, Sub Gerencia de Estudios Económicos, BANCO DE LA REPUBLICA, Bogotá, Colombia.

[3] *Autor Corresponsal:* Investigador Principal, Unidad de Investigaciones, Gerencia Técnica, BANCO DE LA REPUBLICA y Profesor Asociado, Departamento de Estadística, Universidad Nacional de Colombia. Bogotá D. C., Colombia. e-mail:jjulioro@banrep.gov.co.

1 Introduction

Forecasting pervades all human endeavors particularly the design of policies in inflation targeting central banks. In fact, under flexible forecast targeting the most important outputs of the central bank are the short to medium term inflation and economic activity forecasts, which arise from a great variety of sources like the central bank's own suit of models, surveys on external agents and experts, and from internal experts judgment. These forecasts are then comprised into the official ones which serve the board of governors as guide to design its policies. See Svensson (2007), for instance.

The summarization above is based on the assessment of the performance of relevant forecast alternatives, which can be based on two types of evaluation procedures; forecast information disagreement statistics like RMSE, MAE, MAPE, etc., which rank forecast performance depending on the values these statistics assume, and formal forecast performance tests. The former ones are subject to criticism as they lack statistical significance interpretation, and therefore there is marked preference for the latter procedures. See Hyndman and Athanasopoulos (2013) and Diebold and Mariano (1995), for example.

Two statistical tests stand out among the formal methodologies to determine the out of sample performance of a set of $k \geq 2$ forecast procedures of an observable process $\{X_t\}_t$. Consider sample information consisting of k paired sets of size n, $\{X_{i1}, X_{i2}, \ldots, X_{in}\}$ of out of sample forecast errors at a fixed horizon h, arising from $i = 1, 2, \ldots, k$ forecast procedures denoted $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k,$. Given a concave loss function $L(\cdot)$, and letting $\boldsymbol{d}_t = [d_{1t}, d_{2t}, \ldots, d_{k-1,t}]$ be the loss difference process $d_{it} = L(X_{it}) - L(X_{i+1,t})$, Mariano and Preve (2012), MP, propose to test the null of equal expected loss

$$H_0 : E(\boldsymbol{d}_t) = \mathbf{0} \tag{1}$$

against the alternative of at least one difference

$$H_1 : E(\boldsymbol{d}_t) \neq \mathbf{0} \tag{2}$$

using a Wald type statistic $S_{MP} = n\overline{\boldsymbol{d}}'\widehat{\boldsymbol{\Omega}}^{-1}\overline{\boldsymbol{d}}$, where $\widehat{\boldsymbol{\Omega}}$ is an estimate of the asymptotic (long run) variance covariance matrix of $\sqrt{n}\left(\overline{\boldsymbol{d}} - \boldsymbol{\mu}\right)$ and $\overline{\boldsymbol{d}} = \widehat{\boldsymbol{\mu}} = \widehat{E}\left[\boldsymbol{d}_t\right]$[4]. Under the null and the assumptions that the $k \geq 2$ models producing the forecasts are arbitrary and $\boldsymbol{d}_t$ is stationary and has Wold representation, $S_{MP} \xrightarrow{D} \chi^2_{(k)}$ as $n \to \infty$. These authors provide the conditions under which this asymptotic behavior is invariant to permutations of the procedures, and also provide a modified test to correct S_{MP} size in small samples, S_{MPc}.

On the other hand, Giacomini and White (2006), GW, propose two tests; a conditional and an unconditional test for $k = 2$ only. The conditional test determines the

[4] Mariano and Preve (2012) test is a multivariate version of Diebold and Mariano (1995). In this setting it is customary to consider square, absolute, lin-lin and linex losses, among others.

more accurate forecast alternative for a specific horizon while the unconditional tests establishes which of two alternatives was more accurate. The later test coincides with Diebold and Mariano (1995) test for $k = 2$. Under weaker assumptions than MP's, the unconditional test statistic $t_{T,n,h} = \frac{\overline{d}_{T,n}}{\widehat{\sigma}_n/\sqrt{n}}$ has an asymptotic standard normal distribution under H_0 above, for $k = 2$, where T is the sample size used to forecast, $\overline{d}_{T,n} = \frac{1}{n}\sum_{t=1}^{n} d_t$, and σ_n^2 is the long run variance of $\sqrt{n}\left(\overline{d_t} - E[d_t]\right)$.

These tests, MP and unconditional GW, have important advantages over previous methodologies like the possibility of including heterogeneity, time series dependence, and structural breaks within the forecast sample. However, they have important drawbacks as well. First, these tests assess the relative performance of competing forecast alternatives through expected loss order only. By focusing on these figures many features of the forecast error and loss distributions of competing procedures are disregarded, which may lead to sub-optimal choices. Second, their tests statistics have asymptotic rather than exact distributions under the null, which may lead to considerable size distortions in small samples. And third, they have not been proved to be UMP, i.e. Uniformly Most Powerful. Therefore, room for improvement still remains.

We propose a two step methodology to assess the performance of a set of $k \geq 2$ forecast procedures by exploring the distribution of forecast errors and error losses. In the first step forecast alternatives are chosen to have desirable unconditional forecast error distribution properties. In the second, the remaining procedures are ranked according to the stochastic order of error losses. Under the assumption that there is a strict stochastic order among the loss functions of all k procedures, we propose to use the maximum Jonckheere (1954) test statistic among all alternative order permutations to rank their joint performance. However, if strict order happens only among subsets of alternatives, multiple comparisons tests reveal their forecasting ability.

Using Jonckheere's test for our procedures has several advantages over the aforementioned tests as well. First, Jonckheere's test compares the distribution of error losses instead of expected losses. More specifically, the alternative of loss stochastic order is stronger than the alternative of expected loss order, being the former closer to "highly desirable" forecast behavior as it takes into account non systematic forecast error departures from optimal forecasts. Second, Jonckheere's test statistic has exact small sample distribution under the null, which favors its use in many situations, e.g. when moving windows of small size are analyzed as proposed by Fama and MacBeth (1973). Third, Jonckheere's test is non parametric, that is no distributional sample assumption are imposed before hand. Fourth, if strict stochastic order exists only among subsets of the alternative procedures, the multiple comparisons test version provides their ranking as well. Furthermore, since Jonckheere's test works for non paired samples, its scope is wider as paired samples with missing forecasts can be analyzed. And finally, since Jonckheere's test probes the whole distribution rather than just its moments, it may also be useful to assess the performance

Bayesian forecast error distributions.

We compare the power of the joint and multiple comparisons tests based on Jonckheere (1954)'s test with the power of the corresponding joint and multiple comparison versions of Mariano and Preve (2012) test. We implemented two alternative simulation set ups to estimate the power function, the first follows closely Giacomini and White (2006) and the second is an adaptation of Mariano and Preve (2012).

Finally, we provide provide an illustration of the use of our tests in a real world forecasting situation.

The rest of the paper is distributed in 4 sections apart from the introduction. The second discusses the properties of well behaved forecast errors where we favor the use of non parametric tests with exact small sample distribution test statistics under the null. In the third we describe Jonckheere (1954)'s test and the procedures we propose. The fourth contains the comparison of power functions and a real world illustration. The last contains the conclusions and discussion.

2 What makes a good forecast procedure?

The properties of well behaved forecast are well known in the literature. A forecast alternative is loss optimal if it minimizes the expected loss, i.e. the forecast risk, among forecast alternatives. Clearly, this kind of optimality does not necessarily lead to unbiased forecasts, which arise under squared loss, for instance. However, even in this case the distributions of forecast errors and error losses might not necessarily be well behaved.

The case for forecast error distribution unimodality and symmetry relates to non systematic forecast errors. These types of errors occur when the frequency of forecast errors in particular subsets of its support are unexpectedly large, e.g. under multi modality with wide gaps between the modes or under skewness. In the former case a higher than expected frequency of forecast errors will happen around each mode, and since the modes are wide apart a higher than expected frequency of large errors would arise. In the later case, a higher than expected frequency of positive or negative forecast errors may also emerge.

These facts however, do not necessarily affect forecast unbiasedness, a systematic distribution feature. In fact, strong non systematic departures from symmetry, for instance, may lead to forecast bias, but the implication the other way around is not warranted. Therefore, a clear distinction between systematic (e.g. unbiasedness) and non systematic (e.g. multi modality and lack of symmetry) forecast error distribution behaviors comes up, being the later more general and thus preferable than the former.

Under concave loss $L(\cdot)$ and forecast error distribution symmetry and unimodality, the most common measure of forecast "accuracy" relates to expected loss minimization. Let $\{X_{i1}, X_{i2}, \ldots, X_{in}\}$ be k size n sets of paired samples of forecast errors arising from $i = 1, 2, \ldots, k$ forecast procedures denoted $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_{k},$. Forecast alternative $\mathcal{M}_i$ is said to be "more accurate" than $\mathcal{M}_j \neq \mathcal{M}_i$ if $E[L(X_{i,t})] < E[L(X_{j,t})]$, which guides the hypotheses (1) and (2). It is also said that $\mathcal{M}_i$ is the "most accurate" among all k alternatives if it is "more accurate" than any other alternative.

However, loss optimality is not the only way to measure forecast "accuracy'. In fact, expected loss summarizes the behavior of the distribution of forecast errors through its first moment only, which may lead, once again, to non-systematic loss deviations that depend on the shape of the distribution of error losses. Therefore, a stronger measure of "accuracy" can be obtained by comparing loss distributions as a whole.

We propose to test for loss stochastic order instead of expected losses. More formally, we say that $\mathcal{M}_i$ is "strictly more accurate" than $\mathcal{M}_j \neq \mathcal{M}_i$ if $L(X_{i,t}) \prec L(X_{j,t})$ where $\prec$ means $L(X_{i,t})$ is of *strictly smaller stochastic order* than $L(X_{j,t})$, where "strict stochastic order" is defined as

$$P[L(X_{i,t}) > x] < P[L(X_{j,t}) > x] \quad \forall x \in \mathbb{R}, \tag{3}$$

$P[L(X_{i,t}) > x] = 1 - G_i(x)$, and $G_i(x)$ is the cumulative distribution function of the loss applied to the forecast errors of $\mathcal{M}_i$. It is known that $L(X_{i,t}) \prec L(X_{j,t})$ implies $E[L(X_{i,t})] < E[L(X_{j,t})]$ but the opposite implication is not warranted, and that (3) is equivalent to

$$G_j(x) < G_i(x) \quad \forall x \in \mathbb{R} \tag{4}$$

For instance, if $L(\cdot) \geq 0$ and the bulk of the continuous densities g_i and g_j corresponding to G_i and G_j locate around a point $X_0 > 0$, when g_i is higher than g_j around the bulk, and the tail of g_j is greater than the tail of g_i, $\mathcal{M}_i$ has a higher frequency of low losses than $\mathcal{M}_j$, and $\mathcal{M}_j$ has a higher frequency of high losses than $\mathcal{M}_i$. Therefore, $\prec$ optimality of $\mathcal{M}_i$ with respect to $\mathcal{M}_j$ depends on the relative frequency of losses on subsets of the support of L, in sharp contrast with the comparison of expected losses, which are systematic features of loss distributions[5].

Summarizing, we argue that the behavior of forecasts should not only be evaluated through the comparison of systematic features of forecast error distributions (e.g. unbiasedness) and expected error losses (e.g. expected losses), but through the comparison of non systematic distributional features. This way, forecast comparison has to do with

[5] The statistical term "stochastic order" is known to labor economists as first order "stochastic dominance" who apply it to welfare comparison among populations as in Davidson (2008). In this context, several tests for first and higher order, i.e. restricted, stochastic dominance have been developed, which have not been considered in this work since their test statistics have asymptotic rather than exact small sample known distributions under the null. See Davidson and Duclos (2013), for instance.

relative frequencies of forecast errors and error losses on subsets of their corresponding supports rather than on systematic distributional features like moments. Therefore, non systematic features like unimodality, symmetry and loss stochastic order play a key role in forecast evaluation.

Testing the unconditional distribution of forecast errors: In order to test for the desired non systematic behavior of forecast error distributions discussed in 2, available test for distribution symmetry, unimodality and (under square loss) zero location were chosen. For these tasks we prefer non-parametric exact small sample distribution test statistics whenever they are available as they may have superior properties than parametric or asymptotic tests. Our choices are as follows.

1. Mira (2010) proposed a test to detect density lack of symmetry about an unknown measure of location, μ. Under the assumption that $X_1, X_2, \ldots, X_n$ is an i.i.d. sample from a population with c.d.f. $F(x) = F_0(x - \widetilde{\mu})$, the null $H_0 : F_0(x) = 1 - F_0(-x)$ for all $x \in \mathbb{R}$ is rejected when $|\gamma_1(F_n)| \geq \frac{a_n}{n^{1/2}} S_c(\gamma_1, F_n)$, where $\gamma_1 = 2(\overline{X}_2 - X_{s:n})$, $\overline{X}_2$ and $X_{s:n}$ are the sample mean and median, respectively, $a_n \longrightarrow z_{1-\alpha/2}$ as $n \longrightarrow \infty$ with $z_{1-\alpha/2}$ being the $1-\alpha$ percentile of the standard normal distribution, and $S_c^2(\gamma_1, F_n)$ being a weakly consistent estimate of the asymptotic variance of γ_1.

2. In turn, Hartigan and Hartigan (1985) dip test seems to be one of the few options to test for distribution unimodality. This test measures multimodality in a sample by calculating the maximum difference between the empirical distribution function and a unimodal distribution function that minimizes this maximum difference. This maximum difference helps test the null of unimodality against multi modality.

3. Under square loss, optimal forecast error distribution should be located around zero, and therefore Wilcoxon (1945) one sample zero location (mean/median) test, $H_0 : \mu = 0$, under random sampling might be used. The null is rejected when the signed rank sum statistic $W \geq w_{1-\alpha/2,n}$, where $w_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of Wilcoxon (1945) W distribution.

3 A forecast ability test based on stochastic order

3.1 A stochastic order test

Let $(X_{11}, X_{12}, \ldots, X_{1m_1})$,..., $(X_{i1}, X_{i2}, \ldots, X_{im_i})$,..., $(X_{k1}, X_{k2}, \ldots, X_{km_k})$, be k samples of size $m_1, m_2, \ldots, m_i, \ldots, m_k$, randomly drawn from independent populations with arbitrary and continuous cumulative distributions $F_1(x)$, $F_2(x)$, ..., $F_i(x)$, ..., $F_k(x)$ respectively. Jonckheere (1954) proposed a *joint* non parametric exact small sample distribution test for the null hypothesis of distribution equality

$$H_0 : F_1(x) = F_2(x) = \ldots = F_k(x) = F(x), \quad \forall x \in \mathbb{R} \tag{5}$$

against the alternative of a pre-established stochastic order determined by the first subindex, i in X_{ij},

$$H_1 : F_k(x) < F_{k-1}(x) < \ldots < F_1(x), \quad \forall x \in \mathbb{R} \tag{6}$$

which is equivalent to

$$H_1 : X_1 \prec X_2 \prec \ldots \prec X_k$$

The null (5) is rejected whenever $S > s_{1-\alpha}$, where $s_{1-\alpha}$ is the $1-\alpha$ percentile of Kendall (1962) S distribution. Jonckheere (1954, sec. 4) describes the calculation of the exact distribution percentiles.

To calculate Jonckheere's test statistic, let

$$p_{i\alpha_i j\alpha_j} = \begin{cases} 1 & \text{if } X_{i\alpha_i} < X_{j\alpha_j} \\ 0 & \text{if } X_{i\alpha_i} > X_{j\alpha_j} \end{cases}$$

for $i = 1, \ldots, k-1$; $j = 1 + i$; $\alpha_j = 1, \ldots, m_j$,

$$p_{ij} = \sum_{\alpha_i=1}^{m_i} \sum_{\alpha_j=1}^{m_j} p_{i\alpha_i j\alpha_j},$$

and then

$$S = 2\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} p_{ij} - \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} m_i m_j \tag{7}$$

Some care is to be exercised when interpreting the results of this test. When the null is rejected, stochastic order exists for the particular order the alternative was set up. Non rejection of the null, however, can not be interpreted as distribution equality directly as it may also mean that a different stochastic order is present, no stochastic order exists at all, or stochastic order exists only among subsets of the k alternative procedures. Therefore, further exploration might be required if the null is not rejected for a particular order of the k forecast alternatives.

3.2 Maximum and multiple comparison forecast performance tests based on stochastic order

An important drawback of Jonckheere's test is that the stochastic order under the alternative is already known, which is not generally true. Thus, a procedure to uncover the unknown stochastic order is required. Acknowledging that forecast evaluation can be understood as a one way analysis of variance, a test for the existence of joint stochastic order might provide the answer.

We customize Jonckheere's test for this task in the following way. Let $G_1, G_2, \dots, G_k$ be the c.d.f. of the random variables $L(X_1), L(X_2), \dots, L(X_k)$ for a concave loss function $L(\cdot)$. We propose to test the null of distribution equality

$$H_0 : G_1(l) = G_2(l) = \ldots = G_k(l) = G(l), \ \forall l \in \mathbb{R} \tag{8}$$

against the alternative

$$H_1 : G_{i_k}(l) < G_{i_{k-1}}(l) < \ldots < G_{i_1}(l), \text{ for some } (i_1, i_2, \dots, i_k) \in \mathcal{P}, \ \forall l \in \mathbb{R} \tag{9}$$

which is equivalent to test for stochastic order among the k procedures

$$H_{1A} : L(X_{i_1}) \prec L(X_{i_2}) \prec \cdots \prec L(X_{i_k}), \text{ for some } (i_1, i_2, \dots, i_k) \in \mathcal{P}$$

where

$$\mathcal{P} = \{(i_1, i_2, \dots, i_k) : (i_1, i_2, \dots, i_k) \text{ is a permutation of } (1, 2, \dots, k)\} \tag{10}$$

To test these hypotheses we propose the statistic

$$S_{JKMax} = \max_{(i_1, i_2, \dots, i_k) \in \mathcal{P}} S_{JK,(i_1, i_2, \dots, i_k)} \tag{11}$$

where $S_{JK,(i_1, i_2, \dots, i_k)}$ is the Jonckheere's test statistic in equation (7) for the particular permutation $(i_1, i_2, \dots, i_k)$.

Under the assumption that the $k!$ Jonckheere's test statistics $S_{JK,(i_1, i_2, \dots, i_k)}$ are a random sample, the exact distribution under the null becomes $F_{S_{JKMax}} = H_S^{k!}$ where H is Kendall (1962) S distribution above. However, since the random sample assumption might be too strong in this case, further correction will be dealt with in the simulations.

This test provides more information than Mariano and Preve (2012) joint test. In fact, since JKMax test runs over all possible permutations, in case of rejection the permutation corresponding to the maximum Jonckheere's test statistic becomes the more likely joint stochastic order whereas in case of rejection of Mariano and Preve (2012) null, further exploration is required through other types of tests.

However, (9) might be too strong in real life situations, which leads us to assume that stochastic order happens only for subsets of the k forecast alternatives. To uncover which of these k procedures might be of lower stochastic order, we also propose a multiple comparisons test over the k alternatives. This test follows a straightforward application of (5) and (6) for all possible pairs of alternative forecast procedures, correcting the test significance through a Bonferroni procedure.

4 Results

In this section we report power comparisons between Mariano and Preve (2012) and our tests under two settings; (i) $\boldsymbol{d}_t$ follows an independent, e.g diagonal, VAR(1) process as in Giacomini and White (2006, sec. 5.2.1), and (ii) $\boldsymbol{d}_t$ follows a heterogenous serially correlated and correlated MA(q) process as in Mariano and Preve (2012). We customized the the later in such a way that mean losses are increasing and equally spaced under the alternative. In addition to power comparison, we illustrate the use of our procedures on a real world forecast problem.

4.1 Power comparison under GW simulation

Giacomini and White (2006, sec. 5.2.1) proposed an AR(1) model for $\Delta L_{T,t}$:

$$\Delta L_{i,T,t+1} = \mu_i(1-\rho) + \rho \Delta L_{i,T,t} + \varepsilon_{t+1}, \qquad \varepsilon_{i,t+1} \sim N(0,1) \tag{12}$$

for $i = 1, 2, 3, \ldots, k-1$, where T is the sample size used to obtain the n forecasts we analyze. We set $\mu_1 = \kappa$ and $\mu_i = \mu$ for $i = 2, 3, \ldots, k-1$ such that expected losses are increasing and equally spaced,

$$E\left[\boldsymbol{L}_{t+1}\right] = \left[E(L_{i,T,t+1})\right]_{k\times 1} = \left[\kappa, \kappa+\mu, \kappa+2\mu, \ldots, \kappa+(k-1)\mu\right]^T \tag{13}$$

for a suitable constant $\kappa > 0$ we set at $\kappa = 10$.

The size corrected power function of a given test, $K(\mu) = P\left[\text{Reject the null} \mid \mu\right]$, was estimated as the frequency of rejections under alternative values of $\mu \in \{0, 0.05, \ldots, 1\}$ in the following way. Given specific values of $\rho \in \{0, 0.25\}$, $k \in \{2, 3, 4\}$ and $n \in \{12, 30, 100\}$, 100.000 samples were simulated from (12) under the null, $\mu = 0$, and the frequency of rejections was computed. Whenever this frequency differs from the pre-established test size $\alpha = 0.05$, the critical value was corrected so that $\hat{K}(0) = \hat{\alpha} \approx \alpha$ as close as possible, thus equalizing the size of all tests in order to compare their power[6]. Then, 100.000 samples were simulated from (12) under each alternative $\mu \in \{0.05, \ldots, 1\}$, and the size corrected power function was finally estimated as

$$\hat{K}(\mu) = \frac{\#\text{ of rejections}}{100000} \qquad \forall \mu \in \{0, 0.05, \ldots, 1\} \tag{14}$$

[6] Since JK test statistic is discrete, reaching 0.05 is not possible for small n. In this case the tests sizes were corrected as close as possible to 0.05.

4.1.1 The power of joint tests

Figures B.1 to B.3 display the power of the JKMax and MP tests for $k = 2, 3, 4$ respectively. Figure B.1 reveals that these tests are not only unbiased, but also that power increases as the sample size increases, with an important power gain when $n = 100$. In the same way, this figure also shows that power deteriorates as ρ increases to 0.25, showing that both tests, JKMax and MP are equally non-robust to slight forecast error autocorrelation. Moreover, there is no difference whatsoever between the performance of JKMax and MP tests regardless of the fact that JKMax test is stronger and has stronger assumptions than MP.

Figure B.2 reveals a similar picture as B.1 in terms of power unbiasedness and behavior as n and ρ increase. However, an important feature arises by comparing these figures. For $n = 12$ the power of JKMax test is slightly higher than the power of MP test, but this difference vanishes as n increases. Moreover, the power reduction as ρ increases is similar for both tests and JKMax's power is never lower than MP's. Therefore, JKMax seems to be more powerful than MP for small samples.

Figure B.3 depicts a similar behavior of the test as the preceding ones. However, by comparing it to the previous figures, it shows three important features. First, JKMax is more powerful than MP for small and moderately small samples, and this difference increases with k. Second, JKMax's power is slightly higher than MP's power when $n = 30$ and $\rho = 0.25$. And third, the power of both tests increase as k increases.

Summarizing, under the simulation set up of Giacomini and White (2006), JKMax is as powerful as MP test. More specifically, both tests share the same features as it comes to unbiasedness and behavior as n, ρ and k increase. However, JKMax test is more powerful than MP for small samples, $n = 12$, and the power difference increases with k. Finally, JKMax test seems to be slightly more powerful than MP for moderate n, high auto-correlation and a big number of forecast alternatives.

4.1.2 The power of multiple comparisons tests

Figures B.4 and B.5 depict the minimum, average and upper bound power functions of the JK and MP multiple comparisons tests, for $k = 3, 4$ and $n = 12, 30, 100$. Multiple comparisons were carried out by testing the null $H_0 : E[L_i] = E[L_j]$ against the alternative $H_1 : E[L_i] \neq E[L_j]$ in the case of the MP test, and $H_0 : G_{L_i} = G_{L_j}$ against $H_1 : L_i \prec L_j$ for Jonckheere's test, where the ordered pair (i, j) run along all possible values $(i, j) \in \{(i, j) : i < j \quad i, j = 1, 2, \ldots, k\}$. From the power of these tests we calculate the minimum and average powers. The upper bound power, in turn, is calculated as the minimum between the upper bound Bonferroni correction and the maximum attained rejection probability, 1. The minimum, average and maximum powers are identified by the suffixes Min, Avg and Ub respectively in figures B.4 and B.5. This way of presenting multiple comparison power tests relates to Spjotvoll (1972).

Figures B.4 and B.5 share the following features. First, JK related powers are always higher than the corresponding versions of MP tests. Second, the power of these tests increase with n. And third, there is a uniform autocorrelation related power reduction on both tests. Moreover, power increases with k uniformly, but power differences reduce as k increases. These results are similar to the findings in section 4.1.1 except for the fact that JK seems to be uniformly most powerful than MP regardless of the degree of autocorrelation.

4.2 Power comparison under MP simulation

Mariano and Preve (2012) consider the case when the loss difference vector $\mathbf{d}_t = \Delta \boldsymbol{L}_t$ follows the $k-$dimensional MA(q) process with Gaussian noise given by

$$\boldsymbol{d}_t = \boldsymbol{\mu} + \boldsymbol{\epsilon}_t + \sum_{i=1}^{q} \boldsymbol{\Psi}_i \boldsymbol{\epsilon}_{t-1}, \tag{15}$$

where $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = \rho\mathbf{1} - (\rho - 1)\boldsymbol{I}$, where $\mathbf{1}$ and $\boldsymbol{I}$ are $k \times k$ unity and identity matrices, respectively, $0 \leq \rho < 1$, $\boldsymbol{\Psi}_i = \psi^i \boldsymbol{A}$, where $\boldsymbol{A}$ is a $k \times k$ diagonal matrix with diagonal entries $a_{jj} = 1/\sqrt{j}$ for $j = 1, 2, \ldots, k-1$.

We consider the following parameter values in our simulations; $\rho = 0, 0.5, 0.75$, $\psi = 0.5$, $q = 0, 1, 2$ and $k = 2, 3, 4$. In addition, we set up increasing equally spaced mean losses as in equation 13, that is,

$$E\left[\boldsymbol{L}_t\right] = \left[\kappa, \kappa + \mu, \ldots, \kappa + (k-2)\mu, \kappa + (k-1)\mu\right]^T$$

for $\mu = (0, 0.05, \ldots, 1)$ as in section 4.1.

4.2.1 Power of joint tests

Figures B.6 and B.7 display the power of the joint JK and MP tests under. These figures reveal that power differences are very slight and favor the joint MP test for $q = 0, 1$. However, the power of the joint JK test is higher than the power of the joint MP test when $q = 2$, and the power gap increases with k but reduces as correlation increases. Therefore, the joint MP test is slightly better than the joint JK test when the MA order is low, $q = 0, 1$, but interestingly the joint JK test becomes better than the joint MP test when $q = 2$. However, the power gap reduces as correlation increases.

4.2.2 Power of multiple comparisons tests

In the same way, Figures B.8 and B.9 depict the power functions of multiple comparisons tests based on JK and MP tests. These figures might suggest that multiple comparisons tests based on the JK test are more powerful than multiple comparisons tests based on MP. Moreover, the power gap in favor of JK based tests widen as the order of the MA process increases as well as when correlation increases.

4.3 A real world application

4.3.1 Background

When forecasting the medium term inflation rate Banco de la República, the Colombian Central Bank, attaches great importance to the behavior of inflation over the short run. In fact, the bank combines the forecast of two sets of models in the following way. The "best" Short Run inflation Forecast, SRF, is obtained from a suit of small non structural models, and the medium term forecast is obtained from big structural models by constraining their forecast to cross the SRF at the corresponding horizon thus improving the medium term official forecast performance. Therefore, a key input for the medium term central bank inflation forecast in this country is the SRF.

The suit of models used to obtain the SRF contains a series of total and core inflation forecasts that arise from inflation forecasts of CPI sub baskets. These sub baskets include each CPI item separately, the main item groups like food, housing, clothing and miscellaneous, the tradable and non-tradable baskets, as well as the basket of administered price items. From these forecasts several aggregated inflation and core inflation forecasts are built using their corresponding weights. We consider initially $k_0 = 6$ forecast alternatives for the inflation rate, which are denoted as *G6*, *Artes*, *Combik*, *Ave_m*, *TRUNC* and *Agre_ic*, as described in Martínez and González (2014). Since our interest lies on the SRF, we consider $n = 30$ consecutive forecasts at a horizon of $h = 3$ months, where the number 3 appears as a procedure name suffix in the remaining graphs and tables.

4.3.2 Testing the distribution of forecast errors

According to our proposal we explore first the shape of the distribution of forecast errors. This is performed in figure B.10 and Table A.1. Figure B.10 contains the kernel estimate of the forecast error density of the k_0 forecast alternatives described above. From this figure we may see that most of the densities seem to be centered around zero with slight skewness either way, and they look unimodal and seem to have a similar support. However, it can also be observed that *G6_3*'s forecast error density might be off centered and non symmetric, leading to systematic and non systematic positive forecast errors. However, these figures might not be as informative as proper tests, which we show in Table A.1.

Under square loss, $L(X_{it}) = X_{it}^2$, Table A.1 summarize the results of the tests for desirable forecast error distribution behavior. This Table contains the p-values for the zero location, symmetry and unimodality tests described in section 2. At a 5% significance level the null of zero location is rejected by *G6_3*, leading to discard this forecast alternative. The results in this Table show that there is not enough evidence to reject the remaining forecast alternatives, although the small zero location test p-values related to *Artes3* and *Ave_m3* might raise some concern. Therefore, from the initial $k_0 = 6$ forecast alternatives, *G6_3*'s non systematic forecast error behavior leaves us with just $k = 5$ forecast procedures.

To begin the exploration of the remaining forecast alternatives, we summarized in Table A.2 some elementary forecast statistics. In this Table it can be observed that mean forecast errors can be as high as 13 basis points, and RMSE reach up to 39 basis points, with slightly lower MAE's of up to 33 basis points. According to RMSE the best forecast alternative might be *Agre_ic_3* and the worst *Combik_3*. However, according to MAE the best forecast alternative is *Agre_ic_3* and the worst are *Combik_3* and *Ave_m3*. With these statistics at hand, there is not much to say about forecast performance as they lack significance interpretation.

4.3.3 Testing for performance order

We start by exploring the existence of a strict stochastic order among the losses of the $k = 5$ remaining alternatives along with the exploration of expected error loss differences in Table A.3. The panel "Model" in this table shows the more likely stochastic order suggested by the permutation that reached the maximum JK statistic. In this case the best forecast alternative coincides with previous results but the worst does not. However, as the "Test" panel of the table reveals, the null is not rejected, thus this particular order might not be significantly different from any other, e.g. the one suggested by the RMSE and MAE statistics. It can also be observed in this panel that the joint MP test is not rejected, which suggests that there is no sample evidence in favor of any of the procedures having significantly different expected losses than any other.

To explore these results further, we go on testing the multiple comparisons among the $k = 5$ alternative procedures through JK and MP tests in Table A.4. Pairwise comparisons based on JK are one sided, so the upper panel shows all off diagonal p-values corresponding to the alternative $H_1 : L(X_i) \prec L(X_j)$ where (i, j) are the positions in the matrix. Pairwise comparisons based on the MP test, in turn, are two sided and thus only the upper triangular elements of the panel contain the p-values for the alternative $H_1 : E[L(X_i)] \neq E[L(X_j)]$.

At a 5% level the results in Table A.4 show that *Agre_ic_3* has a stochastically lower loss than *Combik_3*, *Ave_m3* and *TRUNC3*. At a 10% level it also shows that *Artes3* has a stochastically lower loss than *Combik_3*, *Ave_m3* and *TRUNC3*. At any of these levels *Agre_ic_3* and *Artes3* loss distribution is not significantly different, and the loss distributions of *Combik_3*, *Ave_m3* and *TRUNC3* are also equal. Therefore at a 10% level two well differentiated groups are identified by multiple comparisons based on JK. However, no difference was found among the procedures according to MP tests.

These findings are summarized in Figure B.12 as it is customarily reported in one way analysis of variance. A line joining two forecast procedures means that the null of equal distribution was not rejected for either alternative stochastic order, $\mathcal{M}_\rangle \prec \mathcal{M}_|$ or $\mathcal{M}_| \prec \mathcal{M}_\rangle$. In addition, a disjoint line among two alternatives means that the null of equal distribution against a particular stochastic order was rejected, lets say $\mathcal{M}_\rangle \prec \mathcal{M}_|$, where forecast procedures are located in such a way that $i < j$. Therefore, Figure B.12 contains the same information as Table A.4.

In order to understand these findings, Figures B.11 and B.13 show the kernel loss density estimate and the kernel loss distribution estimate. Figure B.11 compares the density of *Agre_ic_3* losses with the loss density of the remaining forecast alternatives.From the upper left panel of this Table, it seems to be clear why the distribution of *Agre_ic_3* and *Artes3* are not significantly different. In fact, the bulk of these densities is quite similar and the tails are somewhat similar as well, with a little bump to the right of *Artes3* loss density. It is also clear why *Agre_ic_3* loss is stochastically smaller than the other three alternatives. The bulk of the remaining alternatives is less protruding than the bulk of the density of *Agre_ic_3* loss, which induces a relatively heavier right tail.

Furthermore, Figure B.13 shows the same picture, but this time we can relate these distributions in terms of 6. The highest distribution seems to be that of *Agre_ic_3* loss, followed closely by the distribution of *Artes3*. The other three distributions are not distinguishable but on some part of the tail.

Finally, the results of this exercise show the advantage of JK tests under small samples. Mariano and Preve (2012) related tests did not detect any difference whatsoever among the whole set of forecast alternatives considered. However, JK related tests not only detected significant differences at 5%, but also determined two clearly specified groups of procedures. These groups characterize for having no stochastic order within and a clear stochastic order between them, thus providing a clear performance picture among them.

5 Conclusion

Traditional forecast disagreement statistics like RMSE, MAE, MAPE, sMAPE, MASE, Theil's U, etc., lack statistical significance interpretation, which led statisticians to look for formal statistical tests. Mariano and Preve (2012) and Giacomini and White (2006) took on this task and proposed tests for the null of expected loss equality. However, there is much more to forecast evaluation than just testing for loss moment order. In order to avoid *non systematic errors* the distribution of forecast errors should be symmetric and unimodal, and the distribution of error losses should be the highest among forecast alternatives. Therefore, in addition to testing for forecast error density symmetry and unimodality, we propose to test for *stochastic loss order* rather than loss moment order as usually proposed in previous works, e.g. Mariano and Preve (2012) and Giacomini and White (2006).

By acknowledging the similarities between forecast performance evaluation and *one way analysis of variance*, we propose two test alternatives. The first tests for strict joint stochastic order among all k forecast procedures through the maximum Jonckheere (1954) test statistic over all possible permutations of the k alternatives. Whenever the null hypothesis is rejected, the permutation corresponding to the maximum provides the more likely stochastic order.

However, if the null is not rejected, several different stochastic orders might not be significantly different from each other and thus further exploration is required. We propose to perform this task through *multiple (i.e. pairwise) comparisons*, which provides a clear picture about the performance of the k forecast procedures.

In order to compare our proposals with previous work, simulations were carried out under the settings studied by Giacomini and White (2006) and Mariano and Preve (2012). Under Giacomini and White (2006) loss differences are independent AR(1) processes while under Mariano and Preve (2012) the vector of loss differences is a heterogenous VMA(q) process. We customized mean loss differences to be consistent with increasing equally spaced mean losses as in Giacomini and White (2006).

Power function comparison shows that JK based tests are at leat as powerful as MP based tests, and under particular circumstances are significantly better. More specifically, the joint JKMax tests is more powerful than MP's under the Giacomini and White (2006) set up especially for small samples and high auto correlation. In addition, under the same simulation setup multiple comparison tests based on JK seem to be more powerful than those based on MP regardless of any other parameters. However, power differences reduce with sample size. Moreover, JK related tests dominate uniformly MP related ones under the MP simulation setup. More precisely, JK joint tests are at least as powerful as MP joint tests under MP's setup, and become significantly better in small samples, high order MA loss differences and a high number of forecast procedures to be tested. In the same way, JK multiple comparison tests seem to dominate MP related ones under MP's setup.

Therefore, we conclude that JK based tests are more powerful than MP based ones particularly for small samples and high moving average orders under Mariano and Preve (2012) setting. Under Giacomini and White (2006) settings JK tests are at least as powerful as MP tests, and there is also clear dominance for small samples and high auto correlations. Moreover, the power difference among these tests increases with the number of forecast alternatives considered and reduces with the sample size.

Furthermore, the forecast evaluation illustration of the procedures proposed show that JK related tests are more sensitive than those derived from MP as the later did not detect any difference whatsoever among the forecast alternatives considered. As a matter of fact, JK related tests found significant differences at a 5% level, and at 10% they detected the existence of two clearly differentiated groups of alternatives. These groups characterize for having no stochastic order within and a very clear stochastic order between them, thus providing a clear forecast performance picture of the alternative procedures. This finding is remarkable as well, given that stochastic order is much more stronger than expected loss order. Finally, by acknowledging that forecast evaluation is similar to a one way analysis of variance, results can be nicely reported using multiple comparison graphs as Figure B.12.

References

Davidson, R. (2008). Stochastic dominance. In S. N. Durlauf & L. E. Blume (Eds.), *The new palgrave dictionary of economics.* Basingstoke: Palgrave Macmillan.

Davidson, R., & Duclos, J.-Y. (2013). Testing for restricted stochastic dominance. *Econometric Reviews*, *32*(1), 84-125.

Diebold, F., & Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*, 134-145.

Fama, E., & MacBeth, J. (1973). Risk, return, and equilibrium: empirical test. *Journal of Political Economy*, *81*, 607-636.

Giacomini, R., & White, H. (2006). Test of conditional predictive ability. *Econometrica*, *74*, 1545-1578.

Hartigan, J., & Hartigan, P. (1985). The Dip test of unimodality. *The Annals of Statistics*, *13*(1), 70-84.

Hyndman, R., & Athanasopoulos. (2013). *Forecasting: principles and practice* (1st ed.). Melbourne, Australia: Open-Access textbooks.

Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika*, *41*, 133-145.

Kendall, M. (1962). *Rank correlation method* (3rd ed.). New York, NY: Hafner publishing company.

Mariano, R. S., & Preve, D. (2012). Statistical tests for multiple forecast comparison. *Journal of econometrics*, *169*, 123-130.

Martínez, W., & González, E. (2014). *Forecasting inflation from dissaggregated data: the colombian case.* Forthcomming. (mimeo, Banco de la República)

Mira, A. (2010). Distribution-free test for symmetry based on bonferroni's measure. *Journal of Applied Statistics*, *26*(8), 959-972.

Spjotvoll, E. (1972). Multiple comparison of regression functions. *The Annals of Mathematical Statistics*, *43*(4), 1076-1088.

Svensson, L. (2007). Inflation targeting. *Princeton University CEPS Working Paper*(144).

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80-83.

Appendices

A Tables

B Figures

Table A.1: Testing the shape of forecast errors

	Test		
Model	Zero location	Symmetry	Unimodality
Agre_ic_3	0.1048	0.5744	0.3161
Artes3	0.0636	0.1051	0.6635
Combik_3	0.1403	0.3297	0.5296
Ave_m3	0.0636	0.0967	0.7133
TRUNC3	0.0803	0.0949	0.9044
G6_3	0.0002	0.3310	0.5498

*Source:*Authors' calculations

Table A.2: Mean forecast errors, ME, root mean square errors, RMSE, and mean absolute errors, MAE

Model	ME	RMSE	MAE
Agre_ic_3	-0.09	0.29	0.24
Artes3	-0.13	0.34	0.26
Combik_3	0.11	0.39	0.33
Ave_m3	0.12	0.38	0.33
TRUNC3	0.11	0.37	0.32

*Source:*Authors' calculations

Table A.3: Joint JKMax and MP loss order tests

Model					Test	
1	2	3	4	5	JK	MP
Agre_ic_3	Artes3	Combik_3	Ave_m3	TRUNC3	0.15	0.50

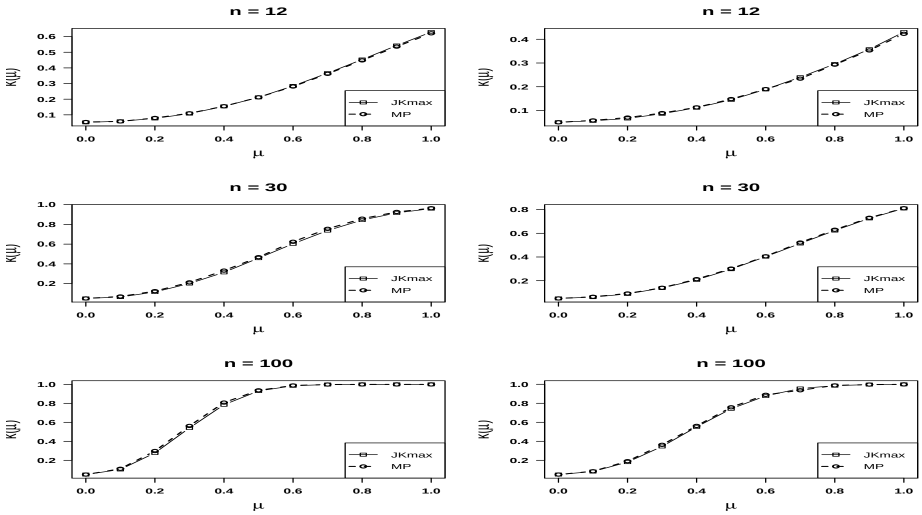
*Source:*Authors' calculations

Table A.4: Multiple comparisons tests

		Model				
Test	Model	Agre_ic_3	Artes3	Combik_3	Ave_m3	TRUNC3
JK	Agre_ic_3		0.45	0.03	0.02	0.02
	Artes3	0.54		0.08	0.06	0.06
	Combik_3	0.97	0.92		0.43	0.43
	Ave_m3	0.98	0.94	0.56		0.49
	TRUNC3	0.98	0.94	0.56	0.50	
MP	Agre_ic_3		0.48	0.13	0.15	0.16
	Artes3			0.60	0.69	0.80
	Combik_3				0.86	0.49
	ave_m3					0.46

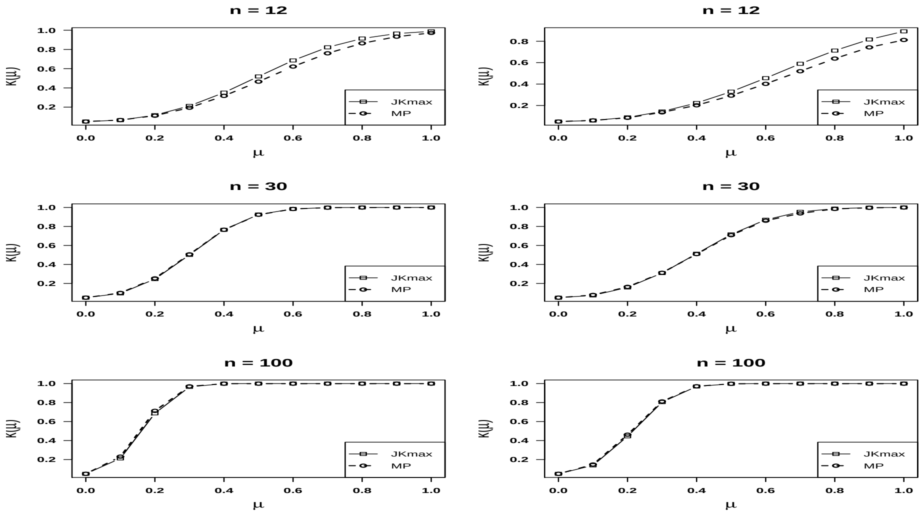
*Source:*Authors' calculations

Figure B.1: Power of JKMax and MP=GW for $k = 2$: $\rho = 0$ (left panel) and $\rho = 0.25$ (right panel)



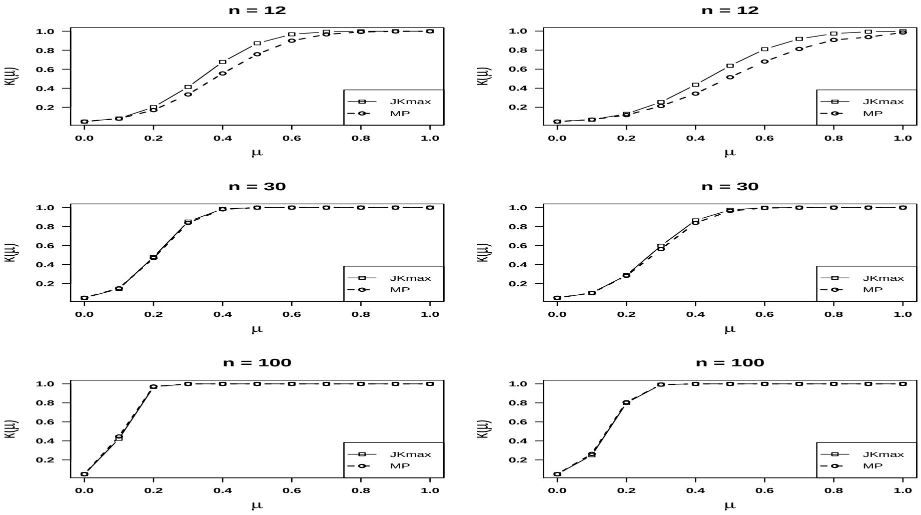
*Source:*Authors' calculations

Figure B.2: Power of JKMax and MP for $k = 3$: $\rho = 0$ (left panel) and $\rho = 0.25$ (right panel)



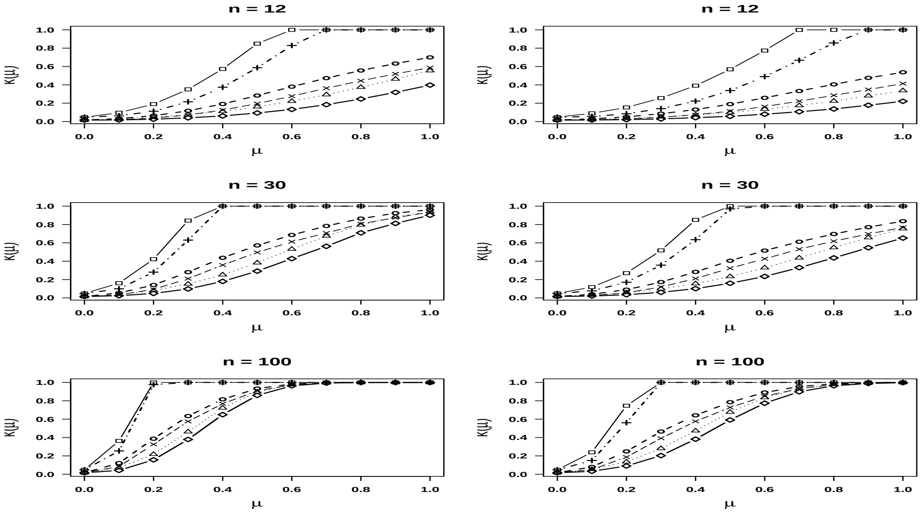
*Source:*Authors' calculations

Figure B.3: Power of JKMax and MP for $k = 4$: $\rho = 0$ (left panel) and $\rho = 0.25$ (right panel)



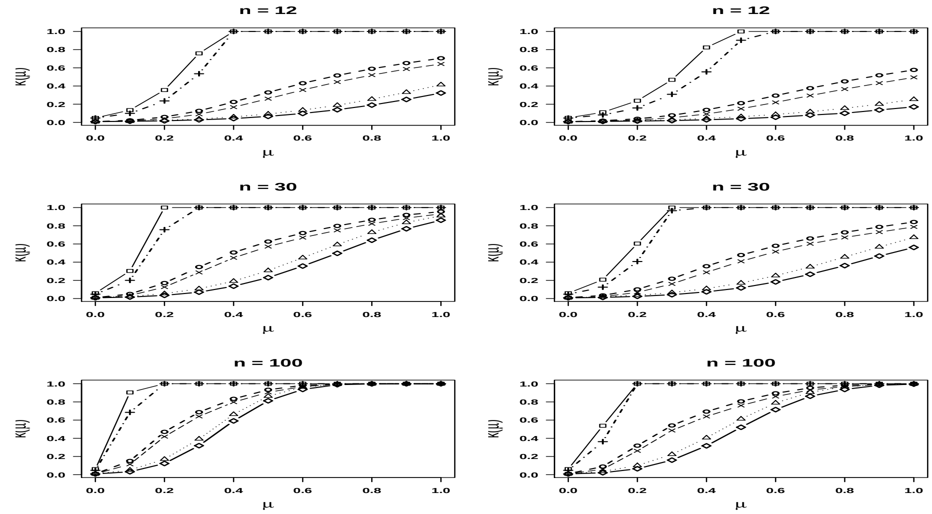
*Source:*Authors' calculations

Figure B.4: Power of JK and MP multiple comparisons tests for $k = 3$: $\rho = 0$ (left panel) and $\rho = 0.25$ (right panel)



◇ MPMin △ JKMin × MPAvg ∘ JKAvg + MPUb □ JKUb
*Source:*Authors' calculations

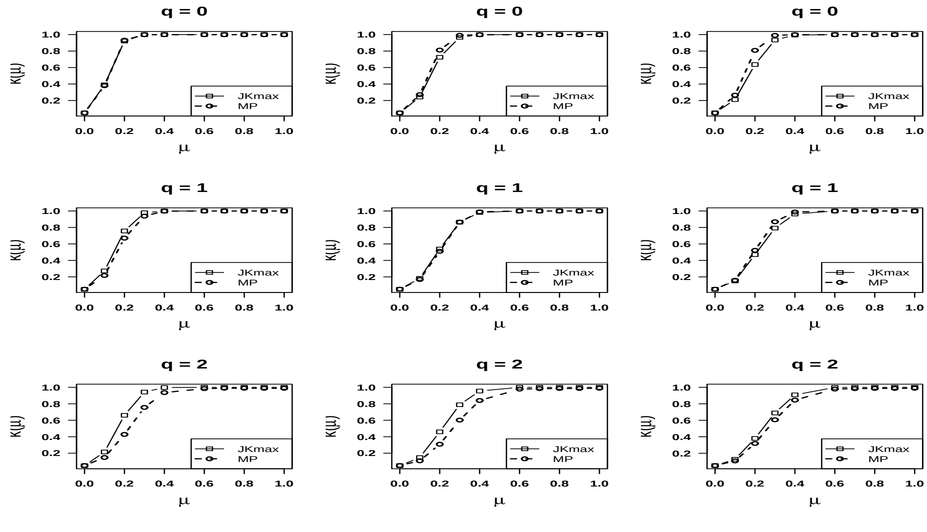
Figure B.5: Power of JK and MP multiple comparisons for $k = 4$: $\rho = 0$ (left panel) and $\rho = 0.25$ (right panel)



◇ MPMin △ JKMin × MPAvg ○ JKAvg + MPUb □ JKUb

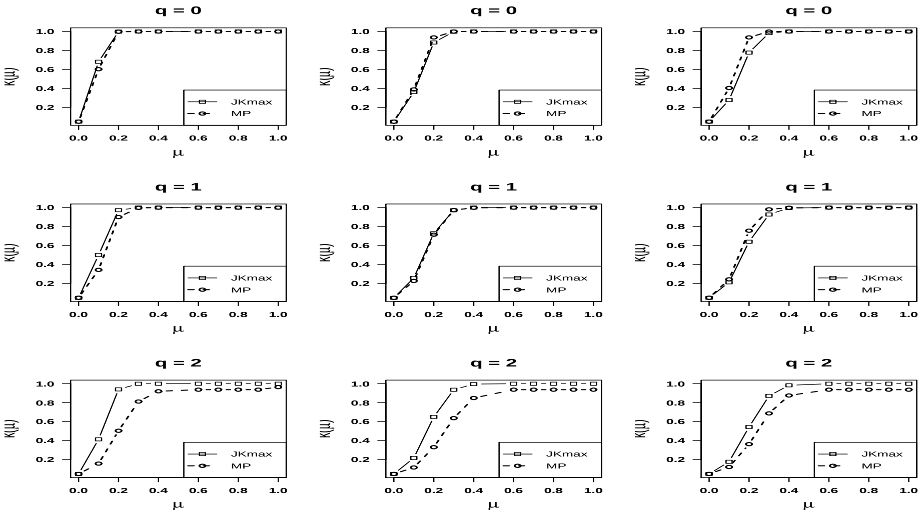
*Source:*Authors' calculations

Figure B.6: Power of JK and MP joint tests for $n = 30$ and $k = 3$: $\rho = 0$ (left panel), $\rho = 0.5$ (middle panel) and $\rho = 0.75$ (right panel) under ordered samples



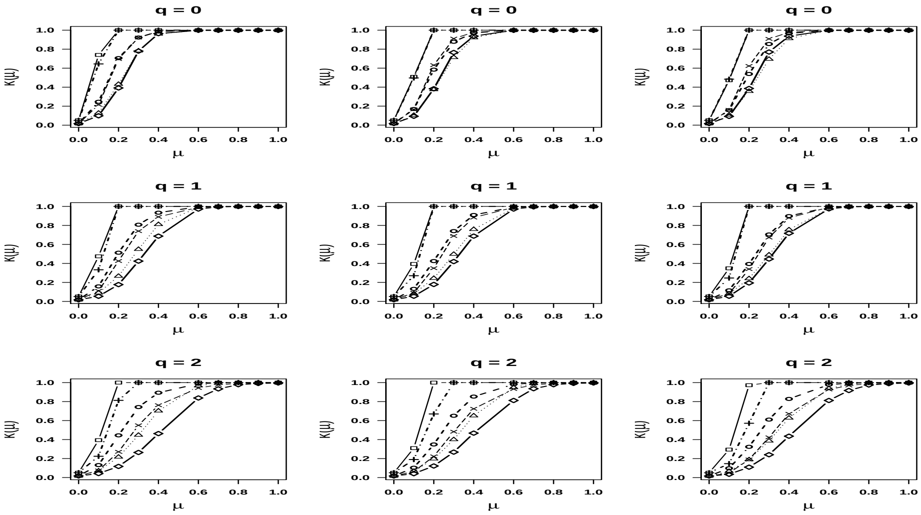
*Source:*Authors' calculations

Figure B.7: Power of JK and MP joint tests for $n = 30$ and $k = 4$: $\rho = 0$ (left panel), $\rho = 0.5$ (middle panel) and $\rho = 0.75$ (right panel) under ordered samples



*Source:*Authors' calculations

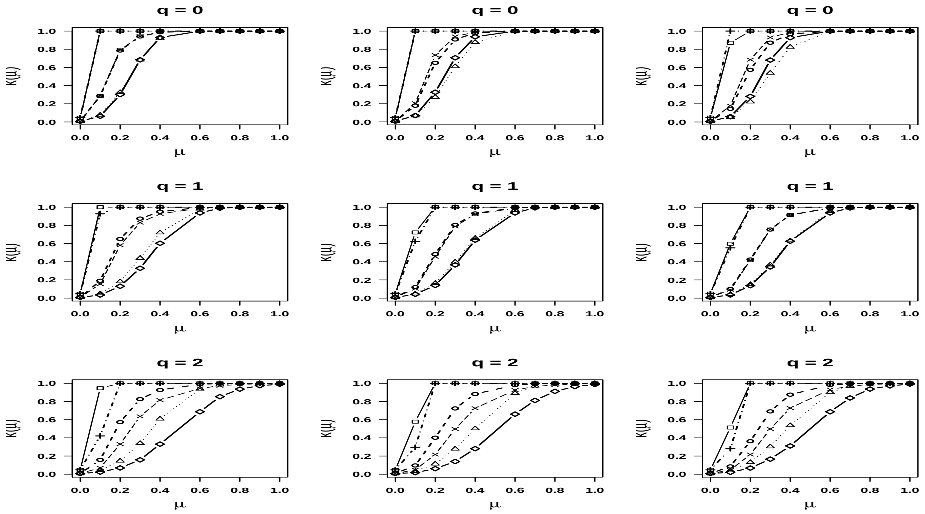
Figure B.8: Power of JK and MP multiple comparisons tests for $n = 30$ and $k = 3$: $\rho = 0$ (left panel), $\rho = 0.5$ (middle panel) and $\rho = 0.75$ (right panel) under ordered samples



◇ MPMin △ JKMin × MPAvg ○ JKAvg + MPUb □ JKUb

*Source:*Authors' calculations

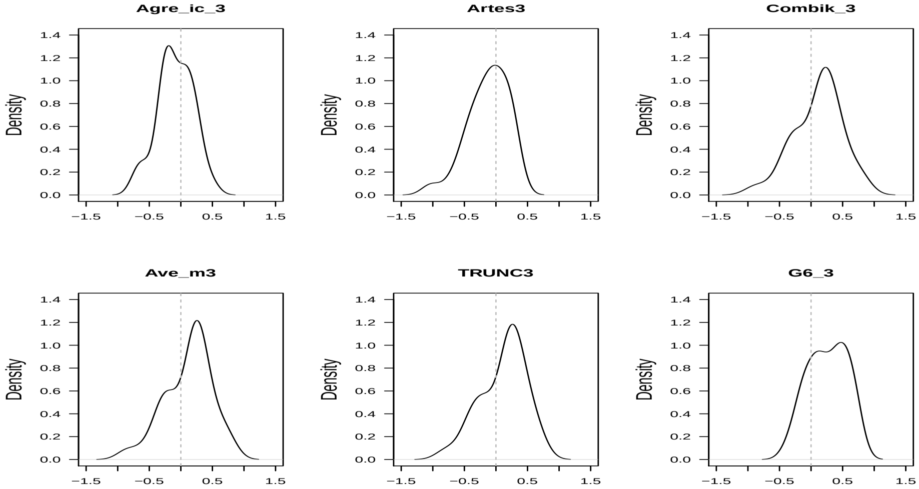
Figure B.9: Power of JK and MP multiple comparisons tests for $n = 30$ and $k = 4$: $\rho = 0$ (left panel), $\rho = 0.5$ (middle panel) and $\rho = 0.75$ (right panel) under ordered samples



◇ MPMin △ JKMin × MPAvg ○ JKAvg + MPUb □ JKUb

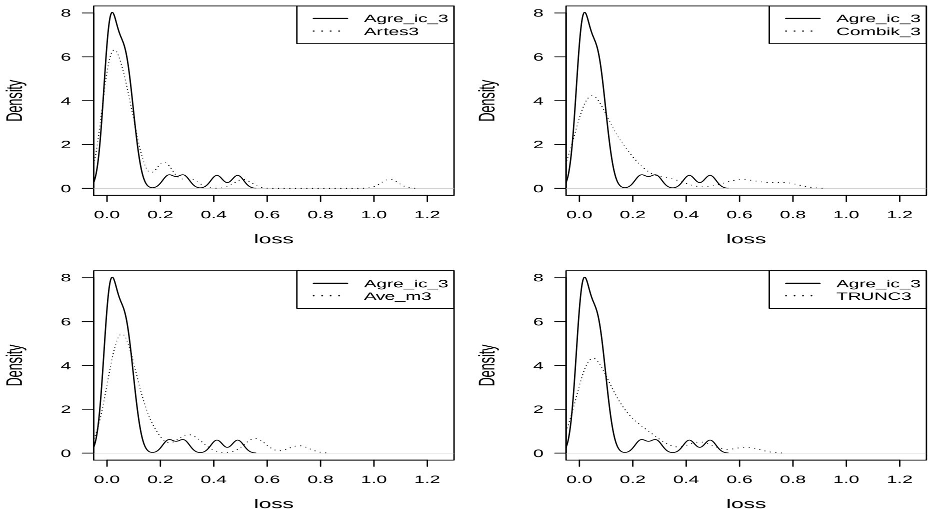
*Source:*Authors' calculations

Figure B.10: Kernel forecast errors density for alternative forecast errors



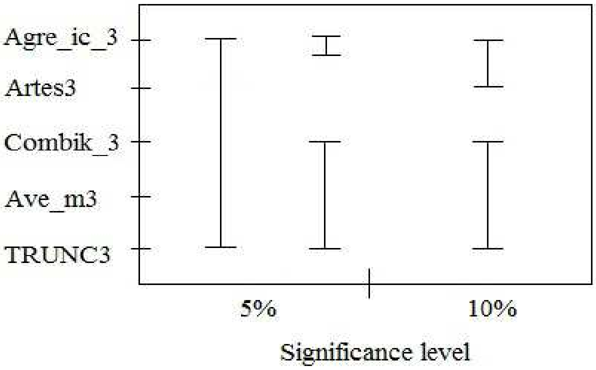
*Source:*Authors' calculations

Figure B.11: Kernel loss density estimate for alternative forecast procedures



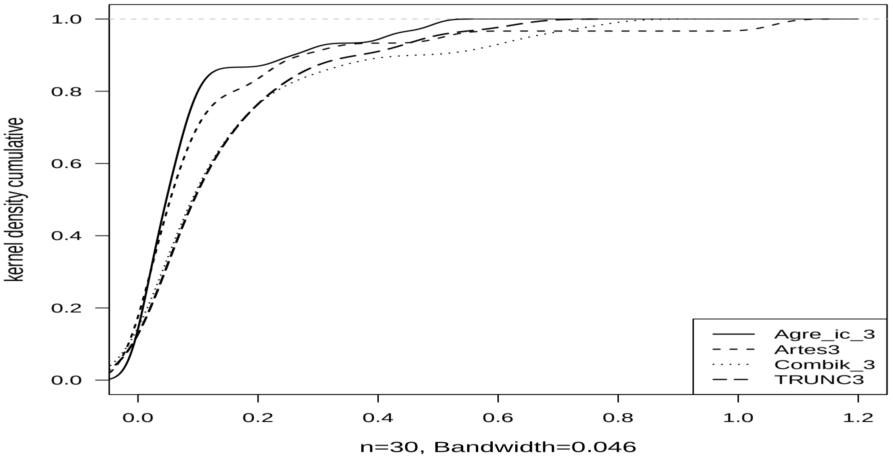
*Source:*Authors' calculations

Figure B.12: Multiple comparison stochastic order results through JK pairwise tests



*Source:*Authors' calculations

Figure B.13: Kernel distribution loss estimate for alternative forecast procedures



*Source:*Authors' calculations