Banco Central de Chile
Documentos de Trabajo

Central Bank of Chile
Working Papers

N° 607

Enero 2011

A BUNCH OF MODELS, A BUNCH OF NULLS AND INFERENCE ABOUT PREDICTIVE ABILITY

Pablo Pincheira

Documentos de Trabajo del Banco Central de Chile
Working Papers of the Central Bank of Chile
Agustinas 1180, Santiago, Chile
Teléfono: (56-2) 3882475; Fax: (56-2) 3882231

Documento de Trabajo
N° 607

Working Paper
N° 607

A BUNCH OF MODELS, A BUNCH OF NULLS AND INFERENCE ABOUT PREDICTIVE ABILITY

Pablo Pincheira
Banco Central de Chile

Abstract

Inference about predictive ability is usually carried out in the form of pairwise comparisons between two competing forecasting methods. Nevertheless, some interesting questions are concerned with families of models and not just with a couple of forecasting strategies. An example of this would be the question about the predictive accuracy of pure time-series models versus models based on economic fundamentals. It is clear that an appropriate answer to this question requires comparing families of models, which may include a number of different forecasting strategies. Another usual approach in the literature consists of comparing the accuracy of a new forecasting method with a natural benchmark. Nevertheless, unless the econometrician is completely sure about the superiority of the benchmark over the rest of the methods available in the literature, he/she may want to compare the accuracy of his/her new forecasting model, and its extensions, against a broader set of methods. In this article we present a simple methodology to test the null hypothesis of equal predictive ability between two families of forecasting methods. Our approach corresponds to a natural extension of the White (2000) reality check in which we allow for the families being compared to be populated by a large number of forecasting methods. We illustrate our testing approach with an empirical application comparing the ability of two families of models to predict headline inflation in Chile, the US, Sweden and Mexico. With this illustration we show that comparing families of models using the usual approach based on pairwise comparisons of the best ex-post performing models in each family, may lead to conclusions that are at odds with those suggested by our approach.

I am are very thankful to Gray Calhoun, an anonymous referee and participants at the 30th International Symposium on Forecasting for their comments and suggestions I also thank Alvaro García for excellent assistance and Gustavo Leyva for programming the stationary bootstrap of Politis and Romano (1994). All errors are mine. Contact Information: pinchei@bcentral.cl , Phone: (56-2) 670 2874, Fax: (56-2) 698 4847. The views expressed in this paper do not necessarily represent those of the Central Bank of Chile or its Board members.

Resumen

La inferencia sobre capacidad predictiva usualmente se realiza a través de comparaciones bilaterales entre dos métodos de proyección en competencia. No obstante, algunas interesantes preguntas se refieren a familias de modelos y no solamente a un par de estrategias de predicción. Un ejemplo de esto sería la pregunta acerca de la capacidad predictiva de los modelos de series de tiempo versus la de modelos basados en fundamentos económicos. Es claro que una respuesta adecuada a esta interrogante involucra la comparación de familias de modelos que pueden incluir un gran número de diferentes estrategias predictivas. Otra práctica usual en la literatura consiste en comparar la precisión de un nuevo método predictivo con un referente natural. No obstante, a menos que el econometrista esté completamente seguro acerca de la superioridad predictiva del referente sobre el resto de los métodos disponibles en la literatura, podría desear comparar la precisión predictiva del nuevo modelo de predicción, y sus extensiones, con un conjunto más amplio de métodos. En este artículo, presentamos una simple metodología para realizar el contraste de la hipótesis nula de igualdad de capacidad predictiva entre dos familias de métodos de proyección. La manera de aproximarnos al problema corresponde a una extensión natural de la confrontación con la realidad que propone White (2000), donde ahora permitimos a las dos familias bajo comparación estar habitadas por un gran número de métodos de predicción. Ilustramos nuestra estrategia de evaluación de hipótesis con una aplicación empírica que compara la capacidad de dos familias de modelos para predecir la inflación del IPC en Chile, Estados Unidos, Suecia y México. Con esta ilustración, mostramos que comparar familias de modelos usando la perspectiva usual basada en comparaciones bilaterales de los pronósticos con mejor comportamiento ex post, puede conducir a conclusiones muy distintas a las que se extraen de nuestra propuesta.

1 Introduction

Forecasts of economic and financial variables are usually important inputs for policy makers in the decision making process. From time to time new forecasting models appear in the literature with the hope of providing a better understanding of the evolution of key economic variables or with the simpler goal of reducing some measure of forecasting error. When evaluating if a novel forecasting approach is useful for prediction, at least three elements are necessary: a measure of accuracy or loss function, a good enough benchmark against which we want to compare the new predictions, and third, an adequate test of predictive ability.

The usual practice, but not the only practice, considers a statistical measure of forecast accuracy like Mean Squared Prediction Error (MSPE), a well known model available in the literature as a benchmark, and tests of equal predictive ability like those developed by Diebold and Mariano (1995) and West (1996). If the object of study are the forecasts themselves rather than the models generating those forecasts, the inference strategy proposed by Giacomini and White (2006) may be preferred.

This usual practice is aimed at comparing the predictive accuracy of two competing forecasts. Even when more than two forecasting methods are considered, often inference is carried out in the form of several pairwise comparisons. In the case of the exchange rate literature, for instance, new models of exchange rate determination are usually compared to the simple random walk model in an attempt to overturn the seminal results in Meese and Rogoff (1983). Nevertheless, some interesting research questions are concerned with families of models and not just with a couple of forecasting strategies. For instance: 1) are time-series models more or less accurate than economic models to predict a given variable? 2) are simple combination strategies more accurate than complex combination schemes to predict a given variable? 3) are forecast that rely solely in the aggregate CPI index more adequate to predict inflation that methods based on disaggregate components? or 4) are linear methods more accurate than non-linear methods? These are all examples of interesting research questions involving the comparison of two families of forecasting methods which may include a number of different forecasting strategies.

In addition, when a new forecasting device is presented in the literature, there is typically some degree of uncertainty surrounding some aspects of this new method. For instance, if a new VAR model is presented, there may be some uncertainty about the number of lags used in each equation, or the number of cointegrating relationships among them. Another good example is the number of different Phillips curve specifications in which the measure of output gap can be calculated in a number of ways and the Phillips curve itself can be augmented with different regressors in several ways as well. Therefore, rather than a new model, the truth is that a family of new models is developed. This family is typically generated by some minor modifications of the main model.

On the other hand, the number of acceptable forecasting methods for traditional economic variables is often large. In this context it is difficult to support the *a priori* selection of a unique benchmark. In the case of inflation the number of well established forecasting models is huge, therefore a more realistic inference approach would be one in which families of models are compared and not just a couple of competing models.

Some interesting contributions dealing with forecasting comparisons including more than two models are the papers by White (2000) and Hansen (2005). Both authors work with a setup in which a number of models are compared to a single benchmark. But what if instead of having a natural benchmark we rather have a family of natural benchmarks? Should we pick our favorite benchmark model and proceed according to White (2000) or Hansen (2005)?

In this paper we address this problem by introducing a natural extension of the approach presented by White (2000) but allowing both families of forecasting devices, the new family and the benchmark family, to be populated by a large number of forecasting methods.

Differing from the results in White (2000), the p-values of our new test need not to be higher than when comparing the best models of both families. This is produced because we are now allowing for specification searches in both families of models: the null and the alternative family. In other words we are accounting for the fact that we could draw a favorable outcome in both of our families just by luck.

The rest of the paper is organized as follows: In Section 2 we introduce our basic econometric setup. In Section 3 we construct an asymptotic test to compare the predictive performance between two families of models. In section 4 we provide an empirical illustration of the use of our test when comparing the predictive ability of two families of inflation forecasts for the case of headline inflation in Chile, Mexico, Sweden and the US. Section 5 concludes and provides a brief summary of our results.

2 Comparing Sets of Forecasting Methods

In this section we consider the following sets of forecasting methods

$$\begin{aligned} M_A &= \left\{\widehat{e}_1^A, \widehat{e}_2^A, ..., \widehat{e}_m^A\right\} \\ M_0 &= \left\{\widehat{e}_1^0, \widehat{e}_2^0, ..., \widehat{e}_J^0\right\} \end{aligned}$$

where $\widehat{e}_i^A$ and $\widehat{e}_j^0$ denote generic one-step-ahead prediction errors from forecasting method i in M_A and forecasting method j in M_0. We call M_A the "alternative family"of forecasting methods and M_0 is called the "null family". We use "hats" in our notation to make explicit the possible dependence of these forecast errors from estimated parameters as in Giacomini and White (2006).

Let us consider a measure of forecast accuracy represented by a generic loss function

$$\begin{aligned} \mathcal{L} &: \mathbb{R}^2 \longrightarrow \mathbb{R} \\ \mathcal{L} &= \mathcal{L}(Y_{t+k}, y_t^p(k)) \end{aligned}$$

where $y_t^p(k)$ is a k-step ahead predictor of Y_{t+k} which uses information available up to time t. Often this loss function can be expressed in terms of an increasing function of the difference between the predictor and the variable it attempts to predict

$$\mathcal{L}(Y_{t+k}, y_t^p(k)) = l(Y_{t+k} - y_t^p(k))$$

the leading example of such a loss function is a quadratic function

$$\mathcal{L}(Y_{t+k}, y_t^p(k)) = l(Y_{t+k} - y_t^p(k)) = (Y_{t+k} - y_t^p(k))^2$$

we will assume we are interested in a loss function that can be expressed as l above.

We consider a null hypothesis according to the unconditional version of the test of predictive ability introduced by Giacomini and White (2006). This is a null expressed in terms of estimates of the parameters of interest. In our case we test

$$H_0 : \mathbb{E}[(l(\widehat{e}_i^A) - l(\widehat{e}_j^0)] = 0 \text{ for all } i = 1, ..., m; \ j = 1, ..., J$$

the alternative hypothesis is one in which there is a forecasting method $\widehat{e}_{i_A}^A$ in family A for which

$$H_A : \mathbb{E}[(l(\widehat{e}_{i_A}^A) - l(\widehat{e}_j^0)] < 0 \text{ for all } j = 1, ..., J$$

that is, we are interested in the identification of a family having the best forecasting method in terms of the loss function l[1].

The next proposition sheds some light regarding the type of statistic we will be using to test our null hypothesis against the previous suggested alternative.

Proposition 1 *The existence of a forecasting method* $\widehat{e}_{i_A}^A$ *in family A for which*

$$H_A : \mathbb{E}[(l(\widehat{e}_{i_A}^A) - l(\widehat{e}_j^0)] < 0 \textit{ for all } j = 1, ..., J$$

is equivalent to the following expression

$$\underset{i \in \{1,2,...,m\}}{Min} \underset{j \in \{1,2,...,J\}}{Max} \mathbb{E}[l(\widehat{e}_i^A)] - \mathbb{E}l(\widehat{e}_j^0)] < 0$$

Proof. See the appendix. ■

It follows that

$$\underset{i \in \{1,2,...,m\}}{Min} \underset{j \in \{1,2,...,J\}}{Max} \mathbb{E}[l(\widehat{e}_i^A) - l(\widehat{e}_j^0)] \tag{1}$$

has a very different behavior under the null and alternative hypotheses. While (1) is exactly zero when the null hypothesis is true, it is strictly negative under the alternative hypothesis.

[1] Notice here that the null and alternative hypotheses do not cover the whole parametric space. This means that we have to be very strict about the interpretation of our test. Rejection of the null hypotesis in the direction of our alternative poses no ambiguity. Nonetheless, no rejection of our null hypothesis cannot be interpreted as its acceptance, as models can be very different in a direction not explored by our test.

3 Building an Asymptotic Test

In this section we construct an asymptotic test based upon the sample analog of (1). For a couple of forecasting methods $\widehat{e}_{it}^{A}$ and $\widehat{e}_{jt}^{0}$ let us define the scalar

$$\overline{X}_t^{(i,j)} \equiv \frac{1}{P}\sum_{t=1}^{P}\left[l(\widehat{e}_{it}^{A}) - l(\widehat{e}_{jt}^{0})\right]$$

where P denotes the sample size of forecast errors.

By considering the difference between $l(\widehat{e}_{it}^{A})$ and every single $l(\widehat{e}_{jt}^{0})$ with $j = 1, ..., J$ we can define the following corresponding column vector

$$\overrightarrow{X}_t^{(i)} \equiv \begin{pmatrix} \frac{1}{P}\sum\limits_{t=1}^{P}\left[l(\widehat{e}_{it}^{A}) - l(\widehat{e}_{1t}^{0})\right] \\ \frac{1}{P}\sum\limits_{t=1}^{P}\left[l(\widehat{e}_{it}^{A}) - l(\widehat{e}_{2t}^{0})\right] \\ . \\ . \\ . \\ \frac{1}{P}\sum\limits_{t=1}^{P}\left[l(\widehat{e}_{it}^{A}) - l(\widehat{e}_{Jt}^{0})\right] \end{pmatrix}$$

Notice that under the null H_0 and mild assumptions, such as those in Giacomini and White (2006), it is possible to show that for each $i = 1, ..., m$

$$\sqrt{P}\frac{1}{P}\sum_{t=1}^{P}\overrightarrow{X_t^{(i)}} = \begin{pmatrix} \sqrt{P}\frac{1}{P}\sum\limits_{t=1}^{P}\left[l(\widehat{e}_{it}^{A}) - l(\widehat{e}_{1t}^{0})\right] \\ \sqrt{P}\frac{1}{P}\sum\limits_{t=1}^{P}\left[l(\widehat{e}_{it}^{A}) - l(\widehat{e}_{2t}^{0})\right] \\ . \\ . \\ . \\ \sqrt{P}\frac{1}{P}\sum\limits_{t=1}^{P}\left[l(\widehat{e}_{it}^{A}) - l(\widehat{e}_{Jt}^{0})\right] \end{pmatrix} \underset{P\to\infty}{\overset{A}{\longrightarrow}} N(0, V_{(J\times J)}^{(i)})$$

with $V_{(J\times J)}^{(i)}$ positive definite. Then, the continuous mapping theorem for convergence in distribution ensures that as P goes to infinity

$$\underset{j\in\{1,2,...,J\}}{Max}\left[\sqrt{P}\frac{1}{P}\sum_{t=1}^{P}\left[l(\widehat{e}_{it}^{A}) - l(\widehat{e}_{jt}^{0})\right]\right] \overset{\mathcal{D}}{\rightarrow}_A \underset{k\ \in\{1,...J\}}{Max}\left\{u_k^{(i)}\right\} \text{ for all } i = 1, ..., m$$

where $\left\{u_1^{(i)}, u_2^{(i)}, ..., u_J^{(i)}\right\}$ is a J-dimensional vector distributed as $N(0, V_{(J\times J)}^{(i)})$. See, for instance, White (2000).

Let us use F_i to denote the following distribution

$$F_i = \underset{k\ \in\{1,...J\}}{Max}\left\{u_k^{(i)}\right\} \text{ for all } i = 1, ..., m$$

it follows that under the null

$$\underset{i\in\{1,2,...,m\}}{Min}\ \underset{j\in\{1,2,...,J\}}{Max}\left[\sqrt{P}\frac{1}{P}\sum_{t=1}^{P}\left[l(\widehat{e}_{it}^{A})-l(\widehat{e}_{jt}^{0})\right]\right]\xrightarrow{\mathcal{D}}_{A}\ \underset{i\in\{1,2,...,m\}}{Min}\{F_1,F_2,...,F_m\}\equiv G \qquad (2)$$

As mentioned in White (2000), when the number of models is small, critical values from G may be obtained using simple Monte Carlo simulations. This can be easily done once consistent estimates of each variance-covariance matrix $V_{(J\times J)}^{(i)}$ are obtained. Otherwise, we can work with bootstrapped critical values. We propose a straightforward generalization of the bootstrap method proposed by White (2000) and also clearly outlined in West (2006). These bootstrapped critical values can be obtained as follows:

1. First, a sequence of P forecast errors for each of the $m \times J$ models is generated using rolling estimation windows.
2. Second, generate B bootstrap samples by sampling with replacement from each original sample. Therefore you end up with a collection of B sequences of P forecast errors for each of the $m \times J$ models. To generate the pseudo-data we use the stationary bootstrap of Politis and Romano (1994).
3. For every possible combination of alternative models i $=1,...,m$ and null models j $=$ $1,...,J$ we compute the bootstrap statistic

$$X_t^{(i,j)*}(b)-\overline{X}_t^{(ij)}\equiv\frac{1}{P}\sum_{t=1}^{P}\left[l(\widehat{e^*}_{it}^{A}(b))-l(\widehat{e^*}_{jt}^{0}(b))\right]-\overline{X}_t^{(ij)}, b=1,2,...,B$$

4. For each $b=1,...,B$ we compute the final statistic

$$\overline{u}_b^*=\underset{i\in\{1,2,...,m\}}{Min}\ \underset{j\in\{1,2,...,J\}}{Max}\sqrt{P}\left[X_t^{(i,j)*}(b)-\overline{X}_t^{(ij)}\right]$$

5. Bootstrapped critical values are finally obtained from the quantiles of the empirical distribution of $\overline{u}_b^*$.

In the next section we illustrate how this procedure works in practice, when comparing two families of inflation forecasting methods.

4 Empirical Illustration

In this section we compare the predictive ability of two different families of forecasting methods. We focus on headline inflation forecasts at different horizons. We consider monthly series of the Consumer Price Index (CPI) for Chile, Mexico, Sweden and the USA. Our sample begins in February 1989 and finishes in December 2008.

We build forecasts for the usual log approximation of year-on-year inflation. In other words we work with π_t^{12} defined as

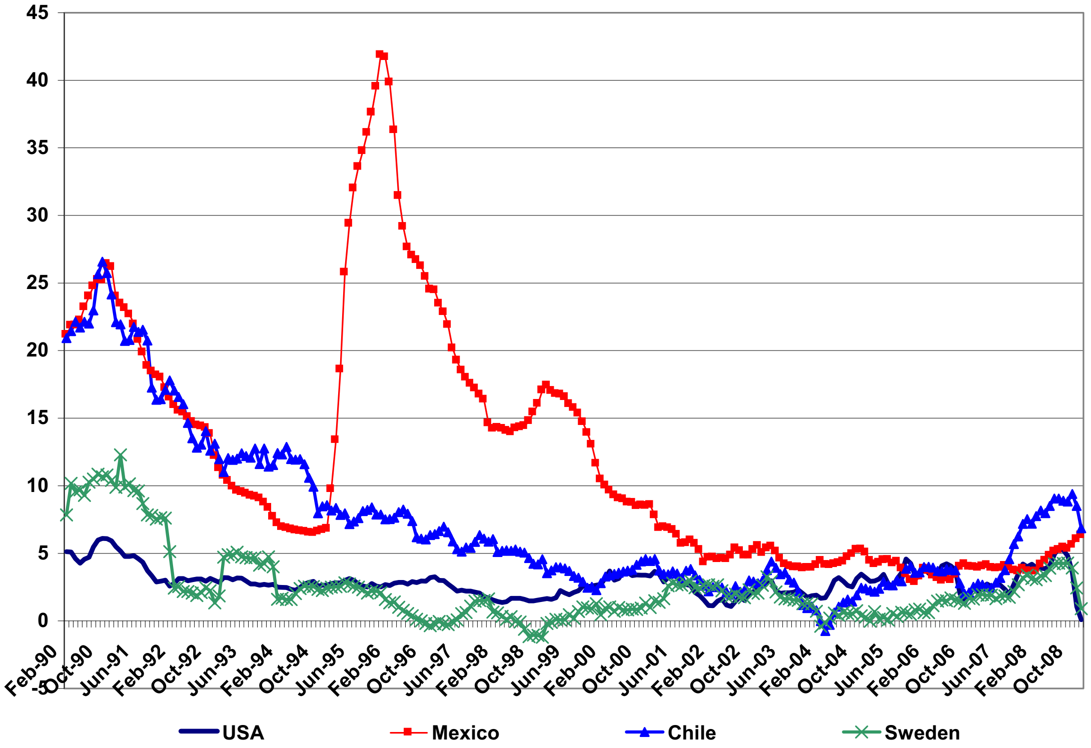
$$\pi_t^{12}=\ln(CPI_t)-\ln(CPI_{t-12})$$

With this transformation our sample reduces to February 1990 - December 2008. We generate sequences of h-step ahead forecasts for every $h = 1, 2, ..., 12$. All forecasts are built from univariate models estimated with rolling windows of 40 observations. Therefore, our first estimation windows spans the period February 1990-May 1993, and our first one-step-ahead forecast is for June 1993. Similarly our first twelve-step-ahead forecast is for May 1994.

Figure 1 shows the evolution of year-on-year inflation for the countries in our sample. It shows a sharp contrast in the cross-country evolution of inflation. Stable inflation is achieved in Mexico during 2002. Chile achieved stable inflation somewhat earlier around 1999. Sweden achieved stable inflation pretty early in our sample, around 1992, whereas the US has shown relatively stable inflation during all our sample period. Because trends in these series may potentially play a role in the construction and evaluation of forecasts, we decide to evaluate the different forecasting methods in a period when most of the countries have achieved stable inflation. This period goes from January 2000 until December 2008. This means that we consider a total of 108 forecasts for every single forecast horizon[2]. In the case of Mexico, we additionally consider another smaller sample starting in July 2002. For this sample we have a total of 78 forecasts at every single horizon. We do this because, as we mentioned earlier, stable inflation in Mexico was achieved not until 2002. Therefore in the subsequent analyses, we will show results for two different samples for Mexico: the usual sample (Mexico) and a shorter sample (Mexico-S).

[2] We consider the same number of forecasts, irrespective of the forecasting horizon, because we have a long series of forecast, that has been truncated to start in January 2000.

Figure 1
Year-on-Year Headline Inflation



In the next subsections we give a brief description of the family of forecasting methods we compare in this empirical exercise.

4.1 Benchmark Methods

The use of different univariate time series models to generate forecasts is fairly usual in the forecasting literature in general, and in the inflation literature in particular. For instance, Atkeson and Ohanian (2001) show that a simple random walk model for year-on-year inflation in the US is very competitive when predicting inflation twelve-months ahead. Giacomini and White (2006), also for the US, present an empirical application in which several CPI forecasts are compared to those generated by a random walk with drift and an autoregression in which the number of lags is selected according to the Bayesian Information Criteria (BIC). Another article using simple univariate benchmarks for the US is Ang, Bekaert and Wei (2007). Among the many methods the authors use, they include an ARMA(1,1) model, a random walk and also an AR(p) model with automatic lag selection according to BIC. Elliot and Timmermann (2008) also explore the ability of several simple univariate models to predict inflation in the US including a simple AR(p) model and single exponential smoothing, which generates the same forecasts as an IMA(1,1) model in which some constraints are imposed over the parameters. More recently, Croushore (2010) also makes use of an IMA(1,1) model as a benchmark when evaluating survey-based inflation forecast for the US. In addition, Stock and Watson (2008) use several different ARMA models as benchmarks to predict inflation in the US. They also use a version of the direct autoregressive model discussed in Stock and Watson (1999). This model looks as follow:

$$\pi^h_{t+h} - \pi_t = \mu^h + \alpha^h(L)\Delta\pi_t + \nu^h_{t+h} \tag{3}$$

where

$$\begin{aligned} \pi^h_{t+h} &= (1200/h)\ln(\frac{CPI_{t+h}}{CPI_t}) \\ \pi_t &= 1200\ln(\frac{CPI_t}{CPI_{t-1}}) \\ \Delta\pi_t &= \pi_t - \pi_{t-1} \end{aligned}$$

and $\alpha^h(L)$ is a polynomial in the lag operator L. Finally, μ^h is just a constant.

Outside of the US the use of univariate time-series models has also become fairly usual. Groen, Kapetanios and Price (2009), for instance, evaluate the accuracy of the Bank of England inflation forecasts using several univariate models, including an AR(p) and the random walk. Similarly, Andersson, Karlsson and Svensson (2007) make use of simple time series models to compare inflation forecasts from the Riksbank. Finally, Pincheira and Alvarez (2009) and Pincheira (2010) also consider ARMA models to construct forecasts for Chilean inflation and GDP growth respectively.

Based on this selective review of the literature and our preliminary exploration, we define the family M_0 as containing the following 11 traditional univariate linear forecasting benchmarks: an AR(1), AR(6), AR(12), ARMA(1,1), ARMA(6,12), ARMA(1,1-12), IMA(1,1), Random Walk, Random Walk with drift, and two versions of the model in (3): The first version selects the lags of the lag polynomio automatically according to AIC, whereas the second version selects these lags according to BIC. Just for clarity the ARMA(1,1-12) is defined as follows:

$$\pi_t = c + \rho\pi_{t-1} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_{12}\varepsilon_{t-12}$$

4.2 Alternative Methods

For the alternative family we rely on the observation of Ghysels *et al* (2006) who mention that the airline model of Box and Jenkins (1970) has a good forecasting performance when predicting seasonal time series. We also rely on early work by Pincheira and García (2009) who show that an extended SARIMA family of models performs well when forecasting Chilean headline inflation at several horizons. This family contains the following eight models

$$\pi_t - \pi_{t-1} = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_{12}\varepsilon_{t-12} \tag{4}$$

$$\pi_t - \pi_{t-1} = \varepsilon_t - \theta_1\varepsilon_{t-1} - \Theta_1\varepsilon_{t-12} + \theta_1\Theta_1\varepsilon_{t-13} \tag{5}$$

$$\pi_t - \pi_{t-1} = \rho(\pi_{t-1} - \pi_{t-2}) + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_{12}\varepsilon_{t-12} \tag{6}$$

$$\pi_t - \pi_{t-1} = \rho(\pi_{t-1} - \pi_{t-2}) + \varepsilon_t - \theta_1\varepsilon_{t-1} - \Theta_1\varepsilon_{t-12} + \theta_1\Theta_1\varepsilon_{t-13} \tag{7}$$

$$\pi_t - \pi_{t-1} = \delta + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_{12}\varepsilon_{t-12} \tag{8}$$

$$\pi_t - \pi_{t-1} = \delta + \varepsilon_t - \theta_1\varepsilon_{t-1} - \Theta_1\varepsilon_{t-12} + \theta_1\Theta_1\varepsilon_{t-13} \tag{9}$$

$$\pi_t - \pi_{t-1} = \delta + \rho(\pi_{t-1} - \pi_{t-2}) + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_{12}\varepsilon_{t-12} \tag{10}$$

$$\pi_t - \pi_{t-1} = \delta + \rho(\pi_{t-1} - \pi_{t-2}) + \varepsilon_t - \theta_1\varepsilon_{t-1} - \Theta_1\varepsilon_{t-12} + \theta_1\Theta_1\varepsilon_{t-13} \tag{11}$$

Interestingly, this extended SARIMA family contains the traditional airline model which is the number (5) above.

The models used by Pincheira and García (2009) display an outstanding predictive performance for Chile when compared to a traditional family of univariate benchmarks similar to that presented in the previous Subsection. It results natural to use the same extended SARIMA family to explore its behavior when predicting inflation in other countries. Nevertheless we complement this extended SARIMA family with four more models. These models are basically the same models (5), (7), (9) and (11) with the only difference that the coefficient associated to the moving average term of order thirteen is not restricted to be equal to $\theta_1\Theta_1$ and now is a free parameter. We do this simply to explore the predictive performance of the models without the restriction mentioned above.

In summary, we use the following six models:

$$\pi_t - \pi_{t-1} = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_{12}\varepsilon_{t-12} \tag{12}$$

$$\pi_t - \pi_{t-1} = \varepsilon_t - \theta_1\varepsilon_{t-1} - \Theta_1\varepsilon_{t-12} + \theta_1\Theta_1\varepsilon_{t-13} \tag{13}$$

$$\pi_t - \pi_{t-1} = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_{12}\varepsilon_{t-12} - \theta_{13}\varepsilon_{t-13} \tag{14}$$

$$\pi_t - \pi_{t-1} = \rho(\pi_{t-1} - \pi_{t-2}) + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_{12}\varepsilon_{t-12} \tag{15}$$

$$\pi_t - \pi_{t-1} = \rho(\pi_{t-1} - \pi_{t-2}) + \varepsilon_t - \theta_1\varepsilon_{t-1} - \Theta_1\varepsilon_{t-12} + \theta_1\Theta_1\varepsilon_{t-13} \tag{16}$$

$$\pi_t - \pi_{t-1} = \rho(\pi_{t-1} - \pi_{t-2}) + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_{12}\varepsilon_{t-12} - \theta_{13}\varepsilon_{t-13} \tag{17}$$

and the same six models plus a drift, which makes a total of twelve models. We will label this alternative family of models as Extended Sarima Family (ESF).

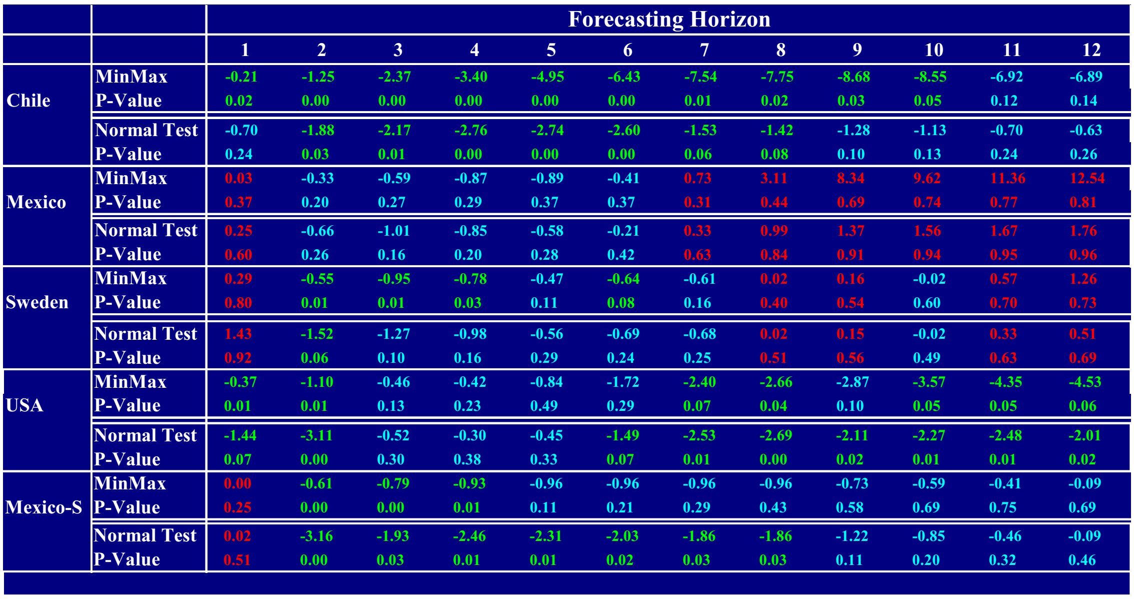
We present the results of our empirical exercise next.

4.3 Empirical Results

Table 1 below shows the results of the MinMax statistic in (2), the traditional core statistic of the Diebold and Mariano (1995) test, that we call in the table "Normal Test", and the resulting p-values associated with both statistics. While the MinMax statistic is comparing the alternative and the benchmark family of models, the "Normal Test" is nothing but the Diebold and Mariano (1995) test when comparing the best performing models in each family. Negative values of the statistics indicate that the alternative Extended Sarima Family outperforms the traditional family of models we are considering here.

We use different colors to highlight qualitatively different results. Figures in green indicate that the Extended Sarima Family works better than the benchmark family and this improvement is statistically significant at the 10% level. Figures in light blue indicate that the Extended Sarima Family is doing better than the traditional family but the difference is not statistically significant at the 10% level. Finally, figures in red indicate that the traditional family of models is doing a better job than the Extended Sarima Family, although this superiority may not be statistically significant at the 10% level.

Table 1
Inference About Predictive Ability When Forecasting Headline Inflation



		Forecasting Horizon											
		1	2	3	4	5	6	7	8	9	10	11	12
Chile	MinMax	-0.21	-1.25	-2.37	-3.40	-4.95	-6.43	-7.54	-7.75	-8.68	-8.55	-6.92	-6.89
	P-Value	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.03	0.05	0.12	0.14
	Normal Test	-0.70	-1.88	-2.17	-2.76	-2.74	-2.60	-1.53	-1.42	-1.28	-1.13	-0.70	-0.63
	P-Value	0.24	0.03	0.01	0.00	0.00	0.00	0.06	0.08	0.10	0.13	0.24	0.26
Mexico	MinMax	0.03	-0.33	-0.59	-0.87	-0.89	-0.41	0.73	3.11	8.34	9.62	11.36	12.54
	P-Value	0.37	0.20	0.27	0.29	0.37	0.37	0.31	0.44	0.69	0.74	0.77	0.81
	Normal Test	0.25	-0.66	-1.01	-0.85	-0.58	-0.21	0.33	0.99	1.37	1.56	1.67	1.76
	P-Value	0.60	0.26	0.16	0.20	0.28	0.42	0.63	0.84	0.91	0.94	0.95	0.96
Sweden	MinMax	0.29	-0.55	-0.95	-0.78	-0.47	-0.64	-0.61	0.02	0.16	-0.02	0.57	1.26
	P-Value	0.80	0.01	0.01	0.03	0.11	0.08	0.16	0.40	0.54	0.60	0.70	0.73
	Normal Test	1.43	-1.52	-1.27	-0.98	-0.56	-0.69	-0.68	0.02	0.15	-0.02	0.33	0.51
	P-Value	0.92	0.06	0.10	0.16	0.29	0.24	0.25	0.51	0.56	0.49	0.63	0.69
USA	MinMax	-0.37	-1.10	-0.46	-0.42	-0.84	-1.72	-2.40	-2.66	-2.87	-3.57	-4.35	-4.53
	P-Value	0.01	0.01	0.13	0.23	0.49	0.29	0.07	0.04	0.10	0.05	0.05	0.06
	Normal Test	-1.44	-3.11	-0.52	-0.30	-0.45	-1.49	-2.53	-2.69	-2.11	-2.27	-2.48	-2.01
	P-Value	0.07	0.00	0.30	0.38	0.33	0.07	0.01	0.00	0.02	0.01	0.01	0.02
Mexico-S	MinMax	0.00	-0.61	-0.79	-0.93	-0.96	-0.96	-0.96	-0.96	-0.73	-0.59	-0.41	-0.09
	P-Value	0.25	0.00	0.00	0.01	0.11	0.21	0.29	0.43	0.58	0.69	0.75	0.69
	Normal Test	0.02	-3.16	-1.93	-2.46	-2.31	-2.03	-1.86	-1.86	-1.22	-0.85	-0.46	-0.09
	P-Value	0.51	0.00	0.03	0.01	0.01	0.02	0.03	0.03	0.11	0.20	0.32	0.46

For clarity of exposition we also present charts displaying both the p-values associated to the MinMax test and the p-values associated to the Diebold and Mariano (1995) test applied to the best performing models in each family. Figures 2-6 show in blue the p-values for the MinMax test when inference is carried out at every single horizon from 1 to 12 months ahead. These graphs also show in red the p-values associated to the Normal test for the same forecasting horizons. The key issue to note here is that both sequences of p-values are different, and sometimes fairly different. This is important, because it indicates that the ex-post selection of the best forecasting model in each family, might not be adequate to compare two families of models when there is uncertainty about the best performing model within each family.

Figure 2 presents p-values for Chile. This graph shows that both sequences of p-values may be extremely different at some forecasting horizons. When forecasting one month ahead, for instance, the Normal Test would indicate that there is no statistically significant difference between the two families of models. The MinMax statistic, however, indicates strong rejection of the null hypothesis of equal predictive ability in favor of the Extended Sarima Family. A similar situation occurs for Mexico when using the smaller sample period and forecasts are made 5 to 8 months ahead. At these forecasting horizons we cannot reject the null hypothesis of equal predictive ability using the Normal Test, and at the same time we would reject this null hypothesis in favor of the Extended Sarima Family when using the MinMax test statistic.

For the US, Sweden and the longer sample for Mexico we are also able to detect differences between the two sets of p-values, but they are more subtle. Interestingly, however, these charts provide another remarkable finding. Differing from the results in White (2000), the p-values coming from the Normal test need not to be lower than those coming from the more comprehensive MinMax test. We can see that this is the case for Chile at every single horizon, and also the case for the rest of the countries in our sample but at different forecasting horizons. The cases of Sweden, the US and Mexico also show that we are not supposed to expect the contrary either, as both curves of p-values cross each other at different points. In summary, we do not detect any particular pattern of dominance of one set of p-values over the other.

Figure 2
P-Values for Chile

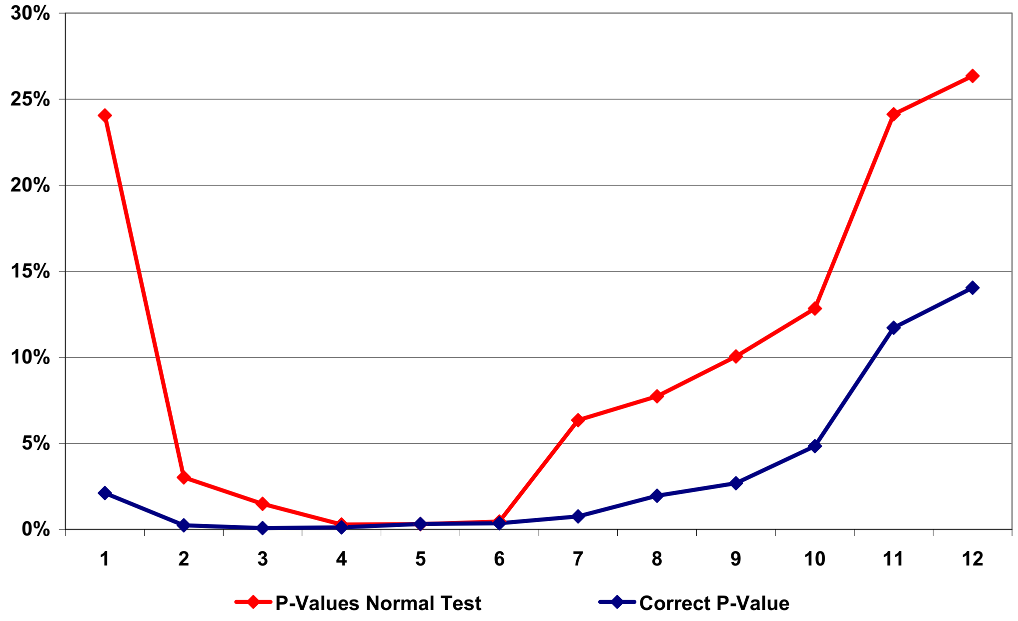


Figure 3
P-Values for Mexico

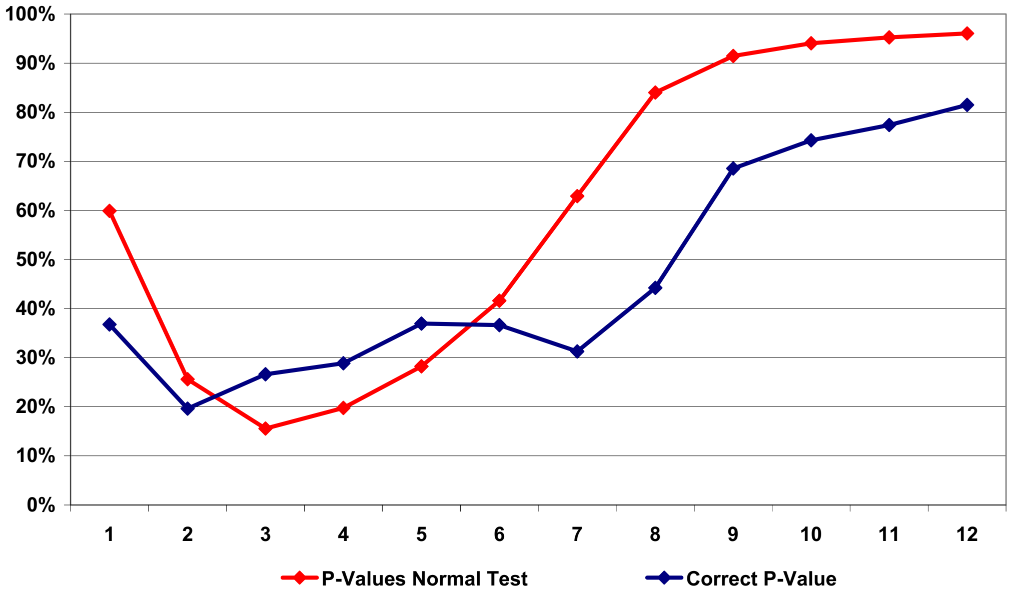


Figure 4
P-Values for Sweden

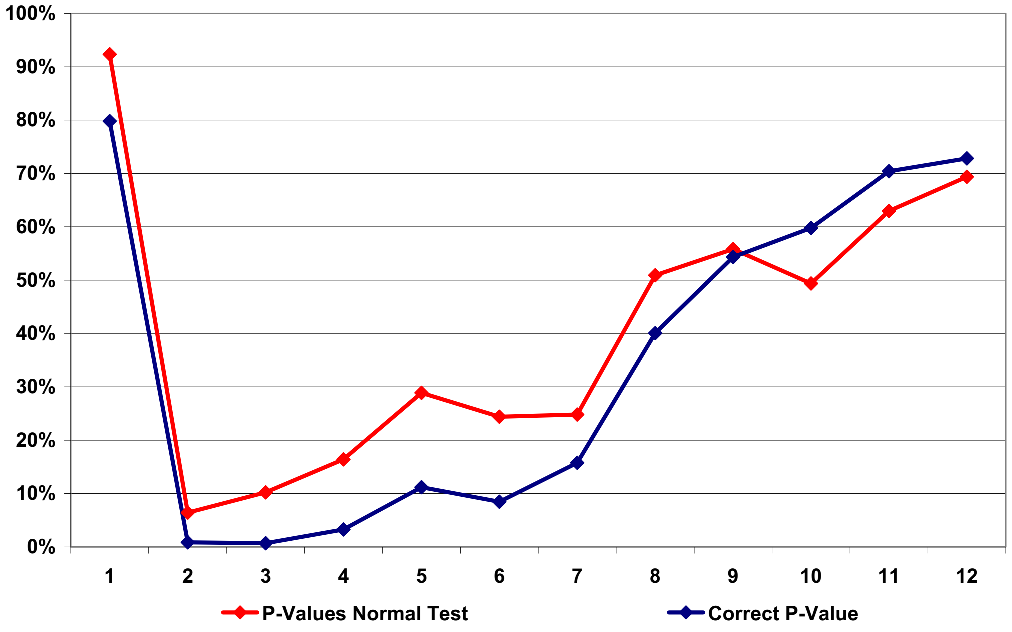


Figure 5
P-Values for the US

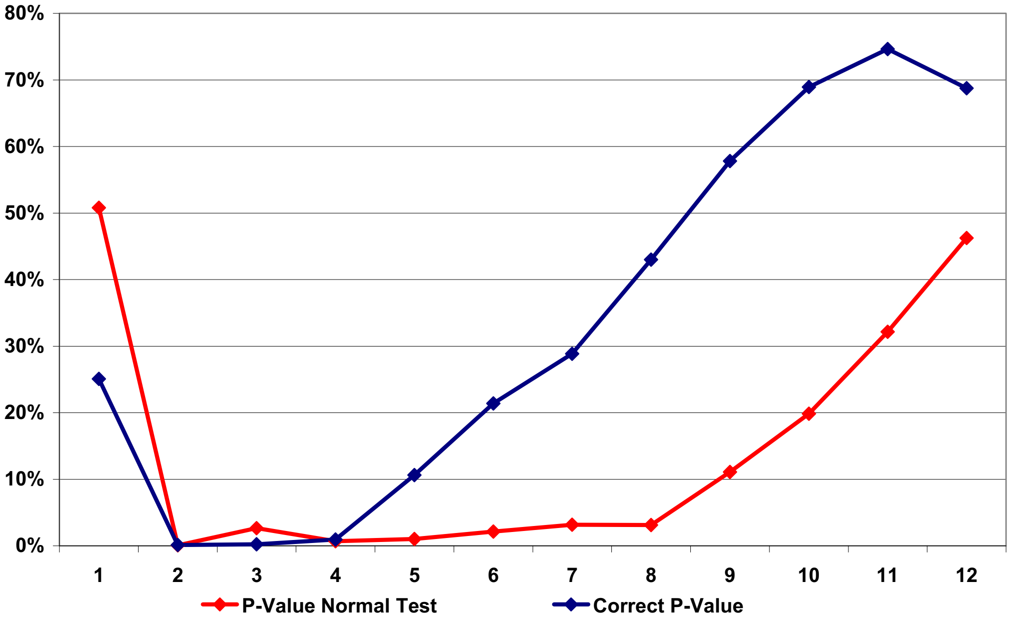
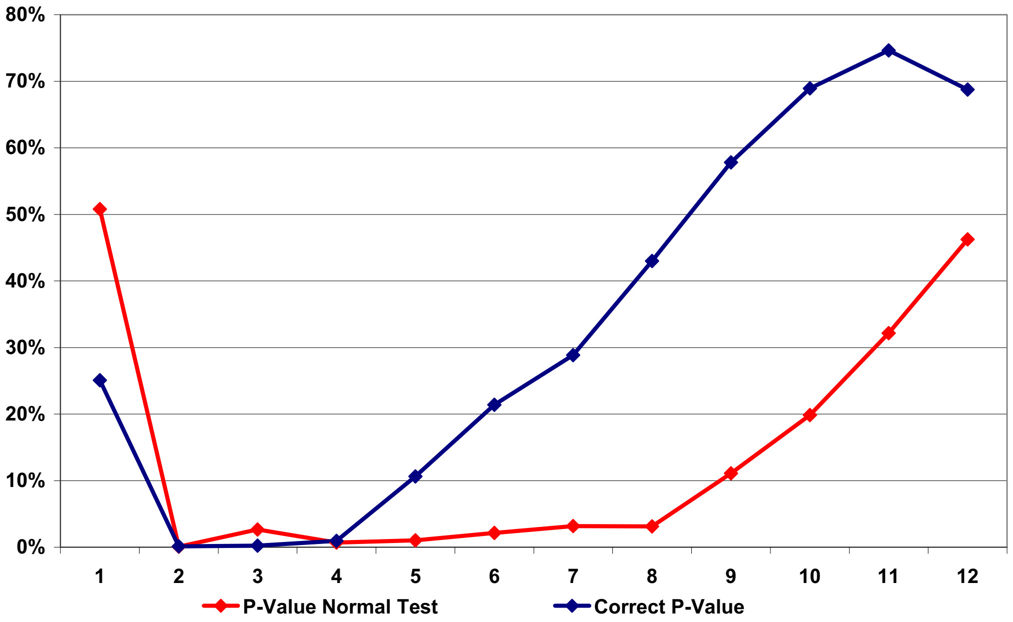


Figure 6
P-Values for Mexico-S

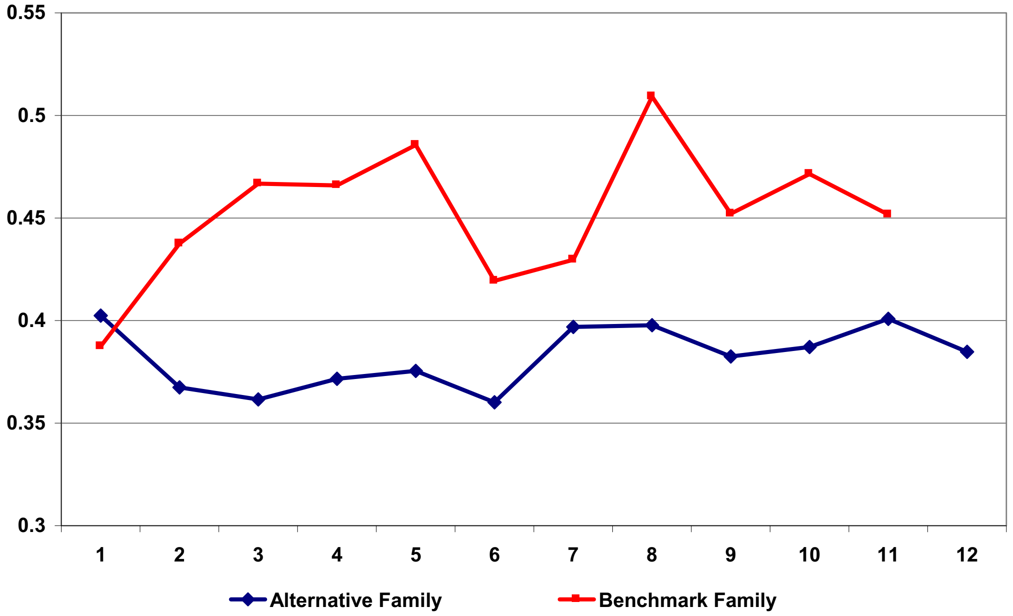


An in-depth understanding of the differences between the two inference strategies we are considering here may be achieved by taking a closer look at Figure 7. This graph depicts the Root Mean Squared Prediction Error (RMSPE) of the twenty three models under consideration when forecasting headline inflation in Chile one-month ahead. The red line shows the RMSPE of the eleven forecasting models in the benchmark family. The blue line shows the RMSPE of every single model in the alternative Extended Sarima Family. Remember from Table 1 and Figure 2 that when forecasting one month ahead, the Normal Test indicates that there is no statistically significant difference between the two families of models. The MinMax statistic, however, indicates strong rejection of the null hypothesis of equal predictive ability in favor of the Extended Sarima Family.

Can we visualize why in this case these two tests provide opposite conclusions? The answer is yes. It turns out that the best performing models in both families display quite similar accuracy. They are both in the range of 0.35 to 0.40. (the actual numbers are 0.387 and 0.360) and basically are pretty close from each other. If we take a look at the other models in both families, however, differences are quite important. The worst performing model from the benchmark family displays a RMSPE exceeding 0.5 which is much higher than the RMSPE of 0.360 corresponding to the best performing model in the Extended Sarima Family. Furthermore, More than 50% of the models in the benchmark family display RMSPE higher than 0.45. On the contrary, most of the models belonging to the Extended Sarima Family show RMSPE below 0.40. All this evidence indicates that, with the exceptions of the best performing models within each family, the two families of models provide quite different results, being more accurate the Extended Sarima Family. This is clearly depicted in Figure 7.

Our interpretation of these results is related to the uncertainty surrounding the identification of the best performing model. If by any chance the researcher has total certainty about the best forecasting models within each family, then she should use this piece of additional information when conducting inference and the Normal test should be employed. On the other hand, if the researcher is not sure about which of the models are the best performers within each family, then she should use the MinMax statistic. By using this statistic the econometrician is implicitly making an acknowledgement of ignorance about the best forecasting model and therefore she is implicitly given positive probabilities to all the models within each family to be the best performers ex post.

Figure 7
RMSPE for Both Families of Models
One-Step-Ahead-Forecasts for Headline Inflation in Chile



5 Conclusion

In this paper we present an extension of the White (2000) reality check approach to develop a testing framework aimed at comparing the predictive ability of two families of forecasting methods. This is an important contribution because many relevant policy and research questions involve the direct comparison of several models and not just of two models. This is because typically when a new forecasting device is presented, there is uncertainty surrounding some aspects of this new method. Therefore rather than a new model, a new contribution generates a family of models in the neighborhood of a central model. A similar situation occurs with the benchmarks available in the literature. In the case of inflation the number of well established and accepted forecasting models is huge. Therefore a more realistic inference approach would be one in which families of models are compared and not just a couple of competing models. Another example relates to different research questions that directly assess the forecasting ability of families of models. This is the case when the researcher wants to know whether linear or nonlinear models predicts better a given economic variable. Similarly, one may be interested in comparing simple and more complex forecasting combination schemes. In the same line one may be interested in comparing the predictive ability of theory-based economic models versus times-series based models. The list of models within each family in this case is huge.

We illustrate the use of our testing framework comparing two families of inflation forecasting methods. The benchmark family consists of a number of simple univariate time-series linear models that traditionally are used in the literature. The alternative family of models is an Extended Sarima Family which includes the famous airline model proposed by Box and Jenkins (1970). This family is an extension of a particular group of SARIMA models, all of which share a unit root and a moving average component of order twelve.

We compare the p-values of our test with those resulting from comparisons of the ex-post best performing models in both families. P-values from these two approaches are in general different and sometimes substantially different. This indicates that when there is uncertainty regarding the best forecasting method within each forecasting family, comparisons of the ex-post best performing strategies within each family may be misleading. Furthermore, and differing from the results in White (2000), the p-values of our new test need not to be higher than when comparing the best models of both families. This is because we are now allowing for specification searches in both families of models: the null and the alternative family. In other words we are accounting for the fact that we could draw a favorable outcome in both of our families just by luck.

A natural extension for future research would consist in compare our results with those of a studentized statistic as suggested by Hansen (2005) and to evaluate the robustness of our test to the presence of irrelevant alternatives.

6 Appendix

6.1 Proof of Proposition 1

The existence of a forecasting method $\widehat{e}^{A}_{i_A}$ in family A for which

$$H_A : \mathbb{E}[(l(\widehat{e}^{A}_{i_A}) - l(\widehat{e}^{0}_{j})] < 0 \text{ for all } j = 1, ..., J$$

is equivalent to the following expression

$$\underset{i\in\{1,2,...,m\}}{Min} \underset{j\in\{1,2,...,J\}}{Max} \mathbb{E}[l(\widehat{e}^{A}_{i})] - \mathbb{E}l(\widehat{e}^{0}_{j})] < 0$$

Proof. Notice that expression

$$\exists\ i_A \in \{1, 2, ..., m\} : \mathbb{E}[(l(\widehat{e}^{A}_{i_A}) - l(\widehat{e}^{0}_{j})] < 0 \text{ for all } j = 1, ..., J$$

is equivalent to

$$\underset{i\in\{1,2,...,m\}}{Min}\mathbb{E}[l(e_i^A)] < \mathbb{E}l(e_j^0)] \text{ for all } j = 1,...,J$$

which is also equivalent to

$$\underset{i\in\{1,2,...,m\}}{Min}\mathbb{E}[l(\widehat{e}_i^A)] < \underset{j\in\{1,2,...,J\}}{Min}\mathbb{E}l(\widehat{e}_j^0)]$$

Notice that

$$\underset{j\in\{1,2,...,J\}}{Min}\mathbb{E}l(\widehat{e}_j^0)] = -\underset{j\in\{1,2,...,J\}}{Max}\left[-\mathbb{E}l(\widehat{e}_j^0)]\right]$$

Therefore

$$\underset{i\in\{1,2,...,m\}}{Min}\mathbb{E}[l(\widehat{e}_i^A)] < -\underset{j\in\{1,2,...,J\}}{Max}\left[-\mathbb{E}l(\widehat{e}_j^0)]\right]$$

which is equivalent to

$$\underset{i\in\{1,2,...,m\}}{Min}\mathbb{E}[l(\widehat{e}_i^A)] + \underset{j\in\{1,2,...,J\}}{Max}\left[-\mathbb{E}l(\widehat{e}_j^0)]\right] < 0$$

which is exactly the same as

$$\underset{i\in\{1,2,...,m\}}{Min}\underset{j\in\{1,2,...,J\}}{Max}\mathbb{E}[l(\widehat{e}_i^A)] - \mathbb{E}l(\widehat{e}_j^0)] < 0$$

■

References

Andersson, M., Karlsson, G. and J. Svensson (2007): The Riksbank's Forecasting Performance. *Economic Review* **3**: 59-75.

Ang, Bekaert and Wei (2007) Do Macro Variables, Assets Markets or Surveys Forecast Inflation Better? *Journal of Monetary Economics* **54** :1163-1212.

Atkeson A. and Ohanian L.E. (2001), Are Phillips Curves Useful for Forecasting Inflation?, *Federal Reserve Bank of Minneapolis Quarterly Review*, **25(1)**: 2–11.

Box G. and Jenkins G. (1970) Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Croushore D. (2010). An Evaluation of Inflation Forecasts from Surveys Using Real-Time Data. *The B.E. Journal of Macroeconomics* **10(1)**: (Contributions), Article 10.

Diebold F and Mariano R. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics* **13:** 253-263.

Elliot G. and Timmerman A. (2008), Economic Forecasting, *Journal of Economic Literature.* **46(1)** :3-56

Ghyles E, D. Osborn and P.M Rodrigues (2006) Forecasting Seasonal Time Series, in Handbook of Economic Forecasting, Volumen 1. G. Elliot, C.Granger and A. Timmermann editors. Elsevier B.V.

Giacomini R. and White H. (2006). Test of Conditional Predictive Ability. *Econometrica* **74**: 1545-1578.

Groen J, Kapetanios G. and S. Price (2009) A real time evaluation of Bank of England forecasts of inflation and growth. *International Journal of Forecasting* **25:** 74-80.

Hansen P. (2005). A Test for Superior Predictive Ability. *Journal of Business & Economic Statistics* **23:** 365-380.

Meese, R. and K. Rogoff (1983). Empirical Exchange Rate Models of the Seventies: Do They Fit Out-of-sample? *Journal of International Economics* **14**:3-24.

Newey, W. K. and K D West (1994). Automatic Lag Selection in Covariance Matrix Estimation, *Review of Economic Studies* **61(4)**:631-53

Newey, W.K. and K.D. West (1987). A Simple, Positive Semidefinite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* **55**: 703-08.

Pincheira P (2010). A real time evaluation of the central bank of Chile GDP growth forecasts. *Money Affairs* Vol XXIII N°**1**, 37-73.

Pincheira P and R. Álvarez (2009). Evaluation of short run inflation forecasts and forecasters in Chile. *Money Affairs* Vol XXII N°**2**,159-180.

Pincheira P. and A. García (2009). Forecasting Inflation in Chile with an Accurate Benchmark. Working Papers Central Bank of Chile N°214.

Politis, D.N. and Romano, J.P.(1994). The Stationary Bootstrap. *Journal of the American Statistical Association* **89:** 1301-1313.

Romano, J.P. and Wolf, M. (2005).Stepwise multiple testing as formalized data snooping. *Econometrica* **73**: 1237-1282.

West K.(1996). Asymptotic Inference About Predictive Ability. *Econometrica* **64:** 1067-1084.

West K. (2006). Forecast Evaluation. In *Handbook of Economic Forecasting*, Elliott G, Granger CWJ, Timmermannn A (eds): 99-134.

White H. (2000). A Reality Check for Data Snooping. *Econometrica* **68**: 1097-1126.

Stock J. and Watson M. (2008). Phillips Curve Inflation Forecasts. *NBER Working Papers* 14322, National Bureau of Economic Research, Inc.

Stock J. and Watson M. (1999). Forecasting Inflation. *Journal of Monetary Economics.* **44:** 293-335.

Documentos de Trabajo
Banco Central de Chile

Working Papers
Central Bank of Chile

NÚMEROS ANTERIORES

PAST ISSUES

La serie de Documentos de Trabajo en versión PDF puede obtenerse gratis en la dirección electrónica: www.bcentral.cl/esp/estpub/estudios/dtbc. Existe la posibilidad de solicitar una copia impresa con un costo de $500 si es dentro de Chile y US$12 si es para fuera de Chile. Las solicitudes se pueden hacer por fax: (56-2) 6702231 o a través de correo electrónico: bcch@bcentral.cl.

Working Papers in PDF format can be downloaded free of charge from: www.bcentral.cl/eng/stdpub/studies/workingpaper. Printed versions can be ordered individually for US$12 per copy (for orders inside Chile the charge is Ch$500.) Orders can be placed by fax: (56-2) 6702231 or e-mail: bcch@bcentral.cl.

DTBC – 606 Enero 2011
College Risk and Return
Gonzalo Castex

DTBC – 605 Enero 2011
Determinants of Export Diversification Around The World: 1962 – 2000
Manuel R. Agosin, Roberto Álvarez y Claudio Bravo-Ortega

DTBC – 604 Enero 2011
A Solution to Fiscal Procyclicality: the Structural Budget Institutions Pioneered by Chile
Jeffrey Frankel

DTBC – 603 Diciembre 2010
Eficiencia Bancaria en Chile: un Enfoque de Frontera de Beneficios
José Luis Carreño, Gino Loyola y Yolanda Portilla

DTBC – 602 Diciembre 2010
Chile’s Structural Fiscal Surplus Rule: A Model – Based Evaluation
Michael Kumhof y Douglas Laxton

DTBC-601 Diciembre 2010
Price Level Targeting and Inflation Targeting: a Review
Sofía Bauducco y Rodrigo Caputo

DTBC-600 Noviembre 2010
Vulnerability, Crisis and Debt Maturity: Do IMF Interventions Shorten the Length of Borrowing?
Diego Saravia

DTBC-599 Noviembre 2010
Is Previous Export Experience Important for New Exports?
Roberto Álvarez, Hasan Faruq y Ricardo A. López

DTBC-598 Noviembre 2010
Accounting for Changes in College Attendance Profile: A Quantitative Life-cycle Analysis
Gonzalo Castex

DTBC-597 Octubre 2010
Fluctuaciones del Tipo de Cambio Real y Transabilidad de Bienes en el Comercio Bilateral Chile - Estados Unidos
Andrés Sagner

DTBC-596 Octubre 2010
Distribucion de Probabilidades Implicita en Opciones Financieras
Luis Ceballos

DTBC-595 Octubre 2010
Extracting GDP signals from the monthly indicator of economic activity: Evidence from Chilean real-time data
Michael Pedersen

DTBC-594 Octubre 2010
Monetary Policy Under Financial Turbulence: An Overview
Luis Felipe Céspedes, Roberto Chang y Diego Saravia

DTBC-593 Septiembre 2010
The Great Recession and the Great Depression: Reflections and Lessons
Barry Eichengreen

DTBC-592 Septiembre 2010
Evidencia de Variabilidad en el Grado de Persistencia de la Política Monetaria para Países con Metas de Inflación
Benjamín García