2. A BUNCH OF MODELS, A BUNCH OF NULLS AND INFERENCE ABOUT PREDICTIVE ABILITY [1]

Pablo PINCHEIRA[2]

Abstract

Inference about predictive ability is usually carried-out in the form of pairwise comparisons between two forecasting methods. Nevertheless, some interesting questions are concerned with families of models and not just with a couple of forecasting strategies. For instance: Are time-series models more accurate than economic models to predict inflation? In this family wise context it is not clear if the methods developed to analyze two models will be useful. We address this problem by presenting a simple methodology to test the null hypothesis of equal predictive ability between two families of forecasting methods. Our approach builds on the reality check presented by White. We illustrate our results comparing the ability of two families of models to predict inflation in Chile, the US, Sweden and Mexico.

Keyword: forecast evaluation, Hypothesis testing, inflation, mean squared prediction error, bootstrap
JEL Classification: C12, C22, C52, C53, E3, E4, F31, F37

I. 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 to compare the new predictions, and third, an adequate test of predictive ability.

[1] *I thank Alvaro García and Carlos Medel for excellent assistance and Gustavo Leyva for programming the stationary bootstrap of Politis and Romano (1994). I also thank Esther Ruiz, Gray Calhoun and an anonymous referee for their comments. All errors are mine.*

[2] *Research Division, Central Bank of Chile. E-mail: ppinchei@bcentral.cl The views expressed in this paper do not necessarily represent those of the Central Bank of Chile or its Board members.*

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 forecasts that rely solely on the aggregate CPI index more adequate to predict inflation than 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 uncertainy 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, instead of 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 original 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 stablished 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.

Different 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. 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 the inference approach to compare the predictive performance between two families of models. In section 3 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 4 concludes and provides a brief summary of our results.

II. Comparing Sets of Forecasting Methods

In this section we consider the following sets of forecasting methods

$$M_A = \{\hat{e}_1^A, \hat{e}_2^A, \ldots, \hat{e}_m^A\}$$

$$M_B = \left\{\hat{e}_1^B, \hat{e}_2^B, \ldots, \hat{e}_j^B\right\}$$

where $\hat{e}_i^A$ and $\hat{e}_i^B$ denote generic one-step-ahead prediction errors from forecasting method i in M_A and forecasting method j in M_B. We call M_A the "alternative family" of forecasting methods, while M_B is called the "benchmark family". We have used "hats" in our notation to make explicit the possible dependence of these forecasts errors from estimated parameters as in Giacomini and White (2006).

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

$$\mathrm{L} : \mathrm{R}^2 \rightarrow \mathrm{R}$$
$$\mathrm{L} = \mathrm{L}(Y_{t+k,} y_t^p(k))$$

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

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

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

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

We assume that 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

$$H_{\mathbf{D}}: E\left[l\left(\hat{e}_i^A\right) - l\left(\hat{e}_j^{\mathbf{D}}\right)\right] = 0 \; for \; all \; \mathbf{i} = \mathbf{1}, \dots, \mathbf{m} \; \mathbf{and} \; j = 1, \dots, J$$

The alternative hypothesis could be one-sided or two-sided. The one-sided version is one in which there is a forecasting method $\hat{e}_{iA}^A$ in family M_A for which

$$H_A: E\left[l\left(\hat{e}_{iA}^A\right) - l\left(\hat{e}_j^{\mathbf{D}}\right)\right] < 0 \; for \; all \;\; j = 1, \dots, J$$

The two-sided version of the alternative hypothesis is one in which there is a forecasting method $\hat{e}_{iA}^A$ in family M_A or a forecasting method $\hat{e}_{j0}^{\mathbf{D}}$ in family $M_{\mathbf{D}}$ for which

$$E\left[l\left(e_{iA}^A\right) - l\left(e_j^{\mathbf{D}}\right)\right] < 0 \; for \; all \; j = 1, \dots, J$$

or

$$E\left[l\left(e_i^A\right) - l\left(e_{j_{\mathbf{D}}}^{\mathbf{D}}\right)\right] > 0 \; for \; all \; i = 1, \dots, m$$

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

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

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

$$H_A: E\left[l\left(\hat{e}_{iA}^A\right) - l\left(\hat{e}_j^{\mathbf{D}}\right)\right] < 0 \; for \; all \;\; j = 1, \dots, J$$

is equivalent to the following expression

$$\underset{\substack{i \in \{1,2,\dots,m\} \\ j \in \{1,2,\dots,J\}}}{Min} \; Max \; E\left[l\left(\hat{e}_i^A\right)\right] - E\left[l\left(\hat{e}_j^0\right)\right] < \mathbf{0}$$

Following the same logic of Proposition 1, it is possible to show that the existence of a forecasting method $\hat{e}_{j\mathbf{D}}^{\mathbf{D}}$ in family $M_{\mathbf{D}}$ for which

$$E\left[l(e_i^A) - l(e_{j_{\mathbf{D}}}^{\mathbf{D}})\right] > 0 \; for \; all \;\; i = 1, \dots, m$$

is equivalent to the following expression

$$\underset{\substack{i \in \{1,2,\dots,m\} \\ j \in \{1,2,\dots,J\}}}{Min} \; Max \; E\left[l\left(\hat{e}_i^A\right)\right] - E\left[l\left(\hat{e}_j^0\right)\right] > \mathbf{0}$$

It follows that

$$\underset{\substack{i \in \{1,2,\dots,m\} \\ j \in \{1,2,\dots,J\}}}{Min} \; Max \; E\left[l\left(\hat{e}_i^A\right)\right] - E\left[l\left(\hat{e}_j^0\right)\right] \qquad (1)$$

has a very different behavior under the null and alternative hypothesis. While (1) is exactly zero when the null hypothesis is true, it is strictly negative under our one-sided alternative hypothesis and strictly different from zero under our two-sided alternative hypothesis.

III. 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 $\overset{\wedge A}{e_{it}}$ and $\overset{\wedge 0}{e_{jt}}$ let us define the scalar

$$\overline{X_t}^{(i,j)} \equiv \frac{1}{P}\sum_{t=1}^{P}\left[l(\overset{\wedge A}{e_{it}}) - l(\overset{\wedge 0}{e_{jt}}) \right]$$

and the corresponding column vector

$$\overrightarrow{X_t}^{(i)} \equiv \begin{pmatrix} \frac{1}{P}\sum_{t=1}^{P}\left[l(\overset{\wedge A}{e_{it}}) - l(\overset{\wedge 0}{e_{1t}}) \right] \\ \frac{1}{P}\sum_{t=1}^{P}\left[l(\overset{\wedge A}{e_{it}}) - l(\overset{\wedge 0}{e_{2t}}) \right] \\ . \\ . \\ . \\ \frac{1}{P}\sum_{t=1}^{P}\left[l(\overset{\wedge A}{e_{it}}) - l(\overset{\wedge 0}{e_{Jt}}) \right] \end{pmatrix}$$

where: P denotes total number of one-step-ahead forecast errors available. 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 $i = 1,...,m$

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

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

$$\underset{j\in\{1,2,\dots,J\}}{Max}\left[\sqrt{P}\frac{1}{P}\sum_{t=1}^{P}\left[l(\hat{e}^{A}_{it})-l(\hat{e}^{0}_{jt})\right]\right]\xrightarrow{D}_{A}\underset{k\in\{1,\dots J\}}{Max}\left\{u^{(i)}_{k}\right\}\text{for all } i=1,\dots,m$$

where: $\left\{u^{(i)}_{1},u^{(i)}_{2},\dots,u^{(i)}_{J}\right\}$ is a J dimensional vector distributed as $N(0,V^{(i)}_{(J\times J)})$. (see White, 2000).

Let us use F_i to denote the following limiting distribution

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

it follows that under the null

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

As mentioned in White (2000), when the number of models is small, critical values of G may be obtained using simple Monte Carlo simulations. This can be easily done once consistent estimates of each variance-covariance matrix $V^{(i)}_{(J\times J)}$ are obtained. Otherwise, we can work with bootstraped critical values. We propose a straightforward generalization of the bootstrap method in White (2000) and also clearly outlined in West (2006). This 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 $=1,\dots,m$ and null models $j=1,\dots,J$ we compute the bootstrap statistic

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

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

$$\overline{u}^{*}_{b}=\underset{i\in\{1,2,\dots,m\}}{Min}\ \underset{j\in\{1,2,\dots,J\}}{Max}\sqrt{P}\left[X^{(i,j)*}_{t}(b)-\overline{X}^{(i,j)}_{t}\right]$$

5. Bootstrap 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.

IV. 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[3]. Our sample begins in January 1959 and finishes in December 2008.

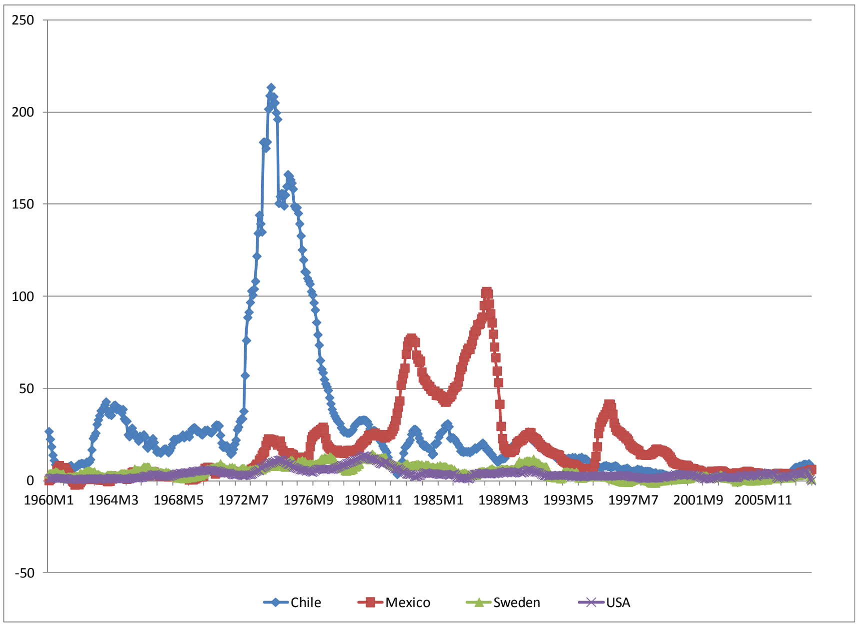
We build forecasts for the usual log approximation of year-on-year inflation. In other words we work with π_t^{12} defined as

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

With this transformation, our sample reduces to January 1960-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 200 observations. Therefore, our first estimation window spans the period January 1960-August 1976, and our first one-step-ahead forecast is for September 1976. Similarly our first 12-step-ahead forecast is for August 1977.

Figure 1

Year-on-Year Headline Inflation



[3] We pick these four countries to illustrate our inference strategy for two main reasons. First, these countries currently belong to the OECD, which ensures data quality and availability. Second, they conform a sample of two stable inflation countries (USA and Sweden) and a sample of two unstable inflation countries (Chile and Mexico), which allows for exploring the behavior of our inference strategy in two different economic environments. We obtained our data from the International Financial Statistics.

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. Chile and Mexico show periods of extremely high and persistent inflation. Compared to them, Sweden and the US showed stable inflation during all our sample period. It is interesting to point out that Chilean inflation reaches a maximum of 214% whereas Mexico reaches a maximum of 103%. We consider the same number of forecasts irrespective of the forecasting horizon just for simplicity. Therefore, we consider a total of P=377 forecasts starting from August 1977 and ending in December 2008.

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

IV.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 inflation12-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 paper 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 forecasts 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:

$$\pi_{t+h}^{h} - \pi_t = \mu^h + \alpha^h(L)\Delta\pi_t + v_{t+h}^{h} \qquad (3)$$

where:

$$\pi_{t+h}^{h} = (1200/h)\ln(\frac{CPI_{t+h}}{CPI_t})$$

$$\pi_t = 1200\ln(\frac{CPI_t}{CPI_{t-1}})$$

$$\Delta\pi_t = \pi_t - \pi_{t-1}$$

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 as benchmarks. 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_D 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 lag of the lag polynomio automatically according to AIC, whereas the second version selects these lags according to BIC. Just for clarity of exposition, the ARMA(1,1-12) is defined as follows:

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

IV.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 (2012), 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 (2012) 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 is 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 (5), (7), (9) and (11) models 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 label this alternative family of models as Extended Sarima Family (ESF). We present the results of our empirical exercise next.

IV.3 Empirical Results

Table 1 below shows the results of the MinMax statistic in (2), the traditional t-statistic of the Diebold-Mariano-West test[4], that we call in the table "Normal Test", and the resulting p-values associated with both statistics for the case in which the alternative hypothesis is one-sided. While the MinMax statistic is comparing the alternative and the benchmark family of models, the "Normal Test" is nothing but the Diebold-Mariano-West test when comparing the best performing models in each family. Negative values of the statistics indicate that the alternative ESF outperforms the traditional family of models we are considering here. Table 2 is similar to Table 1. The only difference is that p-values in Table 2 are calculated for the case in which the alternative hypothesis is two-sided.

We use different colors to highligh qualitatively different results. In Table 1, cells in dark grey with figures in bold indicate that the Extended Sarima Family works better than the benchmark family and this improvement is statistically significant at the 10% level. Figures in normal writing indicate no rejection of the null hypothesis at the 10% level. In Table 2, cells in dark grey have the same meaning as in Table 1. Figures in normal writing indicate that the null hypothesis cannot be rejected at the 10% significance level. Finally, cells in light grey with figures in italics indicate rejection of the null hypothesis at the 10% significance level in favor of the traditional family of models.

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-Mariano-West test applied to the best performing models in each family. To save space we only present charts corresponding to the p-values in Table 2 associated with a two-sided alternative hypothesis. We do this because the difference between our inference

[4] *This test is named after the works of Diebold and Mariano (1995) and West (1996).*

strategy and the the traditional implementation of the Normal test is more noticeable in Table 2[5].

Table 1

Inference About Predictive Ability When Forecasting Inflation
Alternative Hypothesis is One-Sided

Forecasting Horizon													
		1	2	3	4	5	6	7	8	9	10	11	12
Chile	MinMax	-2.22	-9.99	-21.97	-36.44	-5.92	75.96	164.56	288.12	449.44	608.10	777.05	959.64
	P-Value	0.104	0.302	0.689	0.825	0.851	0.898	0.916	0.908	0.918	0.904	0.884	0.864
	Normal Test	-0.80	-1.19	**-1.36**	**-1.38**	-0.11	1.58	2.15	2.28	2.56	2.64	2.57	2.60
	P-Value	0.211	0.118	**0.087**	**0.083**	0.455	0.943	0.984	0.989	0.995	0.996	0.995	0.995
Mexico	MinMax	**-3.80**	**-16.79**	**-41.72**	**-74.85**	**-119.18**	**-177.82**	**-244.06**	**-325.93**	**-454.89**	**-641.39**	**-875.78**	**-722.92**
	P-Value	**0.01**	**0.02**	**0.03**	**0.03**	**0.03**	**0.03**	**0.04**	**0.04**	**0.04**	**0.05**	**0.03**	**0.07**
	Normal Test	**-2.14**	**-1.47**	**-2.03**	**-1.90**	**-2.00**	**-1.70**	**-1.64**	**-1.56**	**-1.52**	-1.08	-1.13	-0.98
	P-Value	**0.016**	**0.071**	**0.021**	**0.029**	**0.023**	**0.045**	**0.050**	**0.060**	**0.065**	0.140	0.130	0.162
Sweden	MinMax	**-0.32**	2.46	3.75	5.38	7.54	10.04	12.76	13.96	16.09	17.58	16.40	30.97
	P-Value	**0.043**	0.783	0.804	0.808	0.850	0.921	0.939	0.915	0.932	0.916	0.865	0.995
	Normal Test	**-2.27**	1.51	1.90	1.93	2.23	2.47	2.30	2.02	1.89	1.68	0.96	2.29
	P-Value	**0.012**	0.935	0.971	0.973	0.987	0.993	0.989	0.978	0.971	0.953	0.832	0.989
USA	MinMax	**-0.24**	0.03	2.06	2.61	3.74	3.86	4.28	5.42	7.15	8.80	12.51	22.99
	P-Value	**0.00**	0.26	0.96	0.97	1.00	0.99	0.98	0.99	0.99	0.99	1.00	1.00
	Normal Test	**-3.52**	0.02	1.40	1.56	1.72	1.63	1.63	1.85	2.07	2.01	2.36	3.93
	P-Value	**0.000**	0.509	0.919	0.940	0.957	0.948	0.948	0.968	0.981	0.978	0.991	1.000

Figures 2-5 show in dotted line 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 solid line 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.

[5] *Very important differences between the p-values of the two strategies under evaluation are achieved only for Chile and Mexico when the alternative hypothesis is one-sided (Table 1). When the alternative hypothesis is two-sided, important differences are achieved for all the countries in our sample.*

Table 2

Inference About Predictive Ability When Forecasting Inflation Alternative Hypothesis is Two-Sided

Forecasting Horizon													
		1	2	3	4	5	6	7	8	9	10	11	12
Chile	MinMax	-2.22	-9.99	-21.97	-36.44	-5.92	75.96	164.56	288.12	449.44	608.10	777.05	959.64
	P-Value	0.208	0.604	0.622	0.350	0.298	0.204	0.168	0.184	0.164	0.192	0.232	0.272
	Normal Test	-0.80	-1.19	-1.36	-1.38	-0.11	1.58	*2.15*	*2.28*	*2.56*	*2.64*	*2.57*	*2.60*
	P-Value	0.423	0.234	0.174	0.167	0.909	0.114	*0.031*	*0.022*	*0.010*	*0.008*	*0.010*	*0.009*
Mexico	MinMax	-3.22	-16.41	-41.72	-74.85	-119.18	-177.82	-244.06	-325.93	-454.89	-623.60	-820.83	-610.33
	P-Value	0.024	0.040	0.052	0.054	0.062	0.056	0.070	0.086	0.082	0.090	0.060	0.130
	Normal Test	-2.41	-2.22	-2.03	-1.90	-2.00	-1.70	-1.64	-1.56	-1.52	-1.51	-1.51	-0.81
	P-Value	0.016	0.027	0.042	0.057	0.045	0.090	0.101	0.120	0.129	0.130	0.131	0.421
Sweden	MinMax	-0.32	2.46	3.75	5.38	7.54	10.04	12.76	13.96	16.09	17.58	16.45	*30.97*
	P-Value	0.086	0.434	0.392	0.384	0.300	0.158	0.122	0.170	0.136	0.168	0.270	*0.010*
	Normal Test	-2.27	1.51	*1.90*	*1.93*	*2.23*	*2.47*	*2.30*	*2.02*	*1.89*	*1.68*	0.96	*2.29*
	P-Value	0.023	0.130	*0.058*	*0.053*	*0.026*	*0.014*	*0.021*	*0.044*	*0.059*	*0.093*	0.337	*0.022*
USA	MinMax	-0.24	0.03	*2.06*	*2.61*	*3.74*	*3.86*	*4.28*	*5.42*	*7.15*	*8.80*	*12.51*	*22.99*
	P-Value	0.004	0.528	*0.082*	*0.060*	*0.010*	*0.018*	*0.042*	*0.028*	*0.020*	*0.016*	*0.010*	*0.000*
	Normal Test	-3.52	0.02	1.40	1.56	*1.72*	1.63	1.63	*1.85*	*2.07*	*2.01*	*2.36*	*3.93*
	P-Value	0.000	0.982	0.161	0.120	*0.085*	0.103	0.103	*0.064*	*0.039*	*0.044*	*0.018*	*0.000*

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 seven months ahead, and at longer horizons as well, the MinMax Test indicates no statistically significant difference between the two families of models at the 10% level. The Normal test, however, indicates strong rejection of the null hypothesis of equal predictive ability in favor of the traditional family of models. This is consistent with the positive values of the Normal Test statistic that are shown in Table 2 at long horizons. A similar but opposite situation occurs for Mexico when forecasts are made 7 to 11 months ahead (see Figure 3). At these forecasting horizons we cannot reject the null hypothesis of equal predictive ability using the Normal Test, and at the same time we reject the null hypothesis in favor of the Extended Sarima family when using the MinMax Test at the 10% level.

For the US and Sweden we are also able to detect differences between the two sets of p-values. The case of Sweden is remarkable, as the Normal test rejects the null hypothesis for most of the horizons in favor of the traditional family of models but, at the same time, the MinMax test rejects the null hypothesis only when forecasting at 1 and 12 months ahead (in favor of the Extended SARIMA Family and the traditional benchmark family, respectively). Finally, the behavior of the two sets of p-values is relatively similar for the US, with the important difference that the MinMax p-values are in general lower than the p-values from the Normal test.

This last point is important. Different from the results in White (2000), figures 2-5 show that the p-values resulted from the Normal test need not to be lower than those resulted from the more comprehensive MinMax test. We already mentioned that this is the case for the US at almost every single horizon, but this also happens for the rest of the countries in our sample at some forecasting horizons. The cases of Sweden, Chile 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 dominance pattern of one set of p-values over the other.

Figure 2

P-Values for Chile

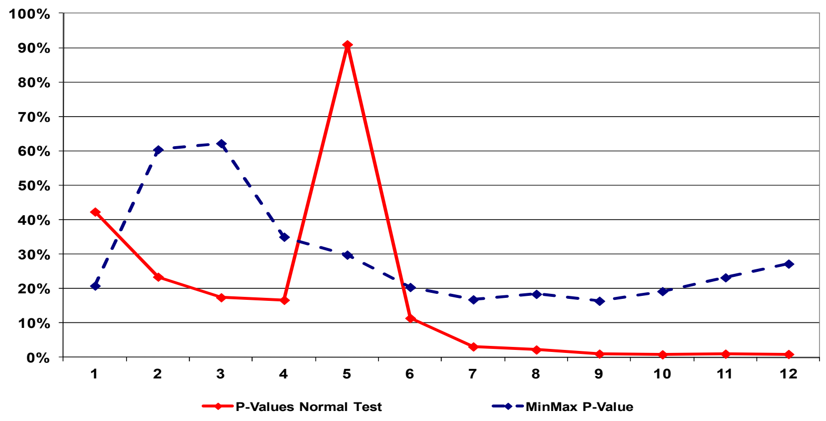


Figure 3

P-Values for Mexico

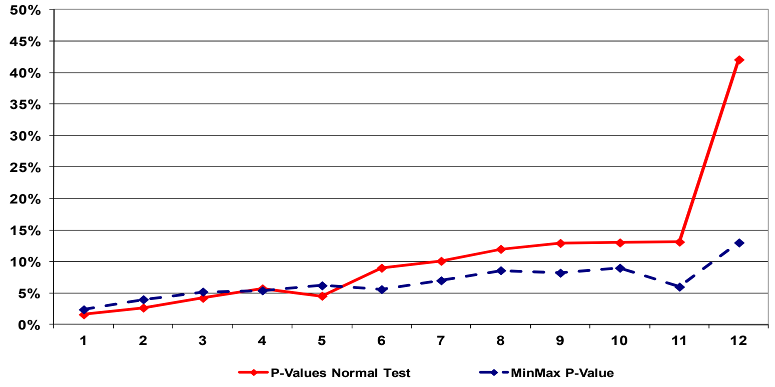


Figure 4

P-Values for Sweden

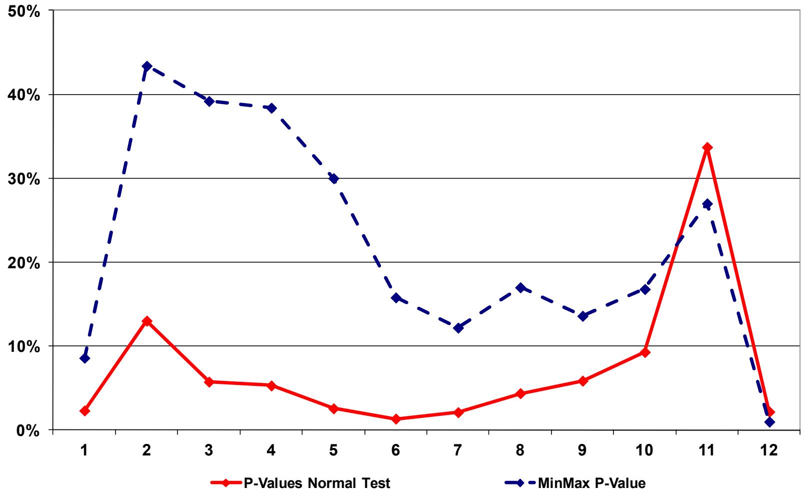
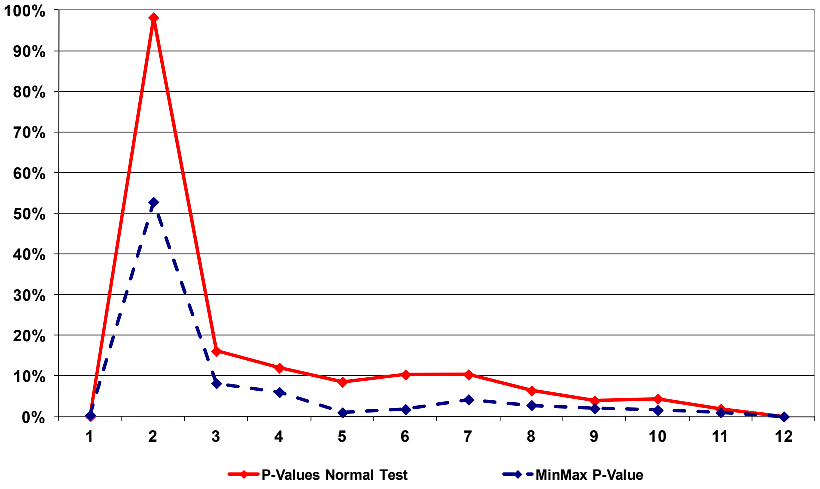


Figure 5

P-Values for the US

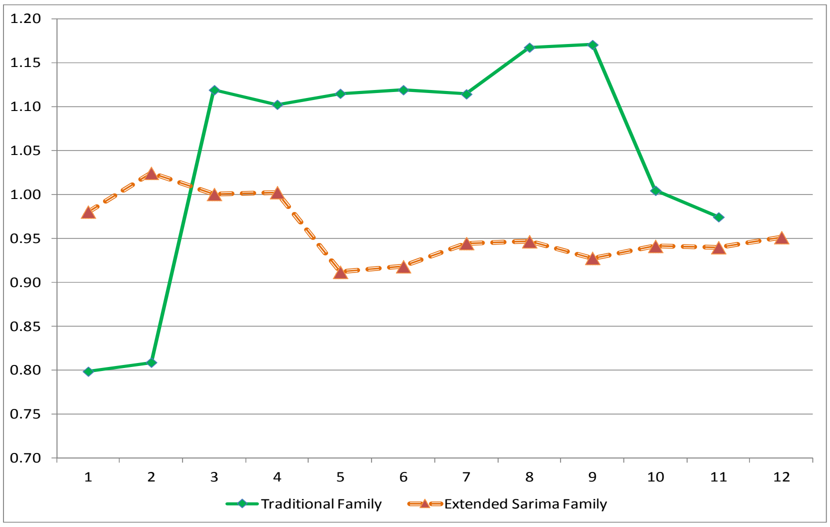


An in-depth understanding of the differences between the two inference strategies we are considering here may be achieved by taking a closer look at Figure 6. This graph depicts the Root Mean Squared Prediction Error (RMSPE) of the twenty three models under consideration when forecasting headline inflation in Sweden three-months ahead. The solid line shows the RMSPE of the eleven forecasting models in the traditional family of benchmarks. The dotted line shows the RMSPE of every single model in the alternative Extended Sarima Family. Remember from Table 2 and Figure

3 that when forecasting three months ahead, the Normal Test indicates rejection of the null hypothesis in favor of the traditional family of models. The MinMax statistic, however, indicates no rejection of the null hypothesis of equal predictive ability at the 10% level.

Figure 6

RMSPE for Both Families of Models
Three-Step-Ahead-Forecasts for Headline Inflation in Sweden



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 different accuracy. The best model in the traditional family displays a RMSPE of 0.8 whereas the best performing model in the Extended Sarima Family shows a RMSPE of only 0.91. This difference is important and the Normal test reflects this fact by rejecting the null hypothesis. If we take a look at the other models in both families, however, differences are quite important in the opposite direction. The worst performing model from the benchmark family displays a RMSPE of 1.17, which is much higher than the RMSPE of 1.024 corresponding to the worst performing model in the Extended Sarima Family. Furthermore, only two models from the traditional family outperform all of the models in the ESF. The third best performing model in the traditional family only beats 4 models in the ESF, whereas the fourth best performing model in the traditional family only beats 2 models from the ESF. All the rest of the 7 traditional models are outperformed by all the models in the ESF. In summary, out of the 132 possible pairwise comparisons between the models in both families, only 30 comparisons favor the traditional benchmarks, whereas 102 favor the Extended Sarima Family.

Our interpretation of these results are 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 he/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 performing models whithin each family, then he/she shoud use the MinMax statistic. By using this statistic the econometrician is implicitly acknowledging ignorance about the best forecasting model. He/she is implicitly given positive probabilities to all the models within each family to be the best performers ex post.

We thing that this acknowledgement of ignorance is relevant as in several occasions it is not simple to pick a best performing model in advance.

V. Conclusions

In this paper, we presented an extension of the White (2000) reality check approach to develop a framework to compare 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 tipically when a new forecasting device is presented, there is uncertainy 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 stablished 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 predict better a given economic variable. Similarly, one may be interested in comparing simple and more complex forecasting combination schemes. In the same line of argument, one may be interested in comparing the predictive ability of theory-based economic models versus times-series based models. The list of families in this case is huge.

By accomodating the test of White (2000) to consider a family of benchmark models we are able to provide a framework for the comparison of two families of models.

We illustrate the use of our statistic comparing two families of inflation forecasting methods. The benchmark family consists of a number of simple univariate time-series linear models that traditionaly are used in the literature to predict inflation. The alternative family of models is an extended SARIMA set of models, 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 are characterized by modeling year-on-year monthly headline inflation with a unit root and with 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 quite 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 different 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 happens because we are now allowing for specification searches in both families of models. 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 reaserch may include the comparisons of our results with those of an studentized statistic, as suggested by Hansen (2005) and the evaluation of the robustness of our test in the presence of irrelevant alternatives.

References

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

Ang, A. Bekaert, G. and M. Wei, 2007. Do Macro Variables, Asset Markets, or Surveys Forecast Inflation Better?. *Journal of Monetary Economics,* 54(4), pp.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), pp. 2-11.

Box G. and Jenkins G. 1970. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisc,o.

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

Diebold F. and Mariano R, 1995. Comparing Predictive Accuracy. *Journal of Business and Economic Statistics,* 13, pp.253-263.

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

Ghyles E, D. Osborn and P.M Rodrigues, 2006. Forecasting Seasonal Time Series. In: G. Elliot, C. Granger and A. Timmermann, eds. *Handbook of Economic Forecasting*, Volume 1. Elsevier B.V.

Giacomini R. and White H, 2006. Test of Conditional Predictive Ability. *Econometrica,* 74, pp.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, pp.74-80.

Hansen P.R. 2005. A Test of Superior Predictive Ability. *Journal of Business&Economic Statistics,* 23, pp.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, pp.3-24.

Pincheira, P, 2010. A real time evaluation of the central bank of Chile GDP growth forecasts. *Money Affairs,* 23(1), pp.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 Á. García, 2012. En busca de un buen marco de referencia predictivo para la inflación en Chile. *El Trimestre Económico*, Fondo de Cultura Económica, 313, pp.85-123.

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

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(5), pp.1098-1126.

Stock, J. and Watson, M., 2008. Phillips Curve Inflation Forecasts, NBER Working Papers 14322,.

Stock, J. and Watson, M., 1999. Forecasting Inflation, *Journal of Monetary Economics*, 44, pp. 293-335.