The Accuracy of the Survey of Professional Forecasters for the Euro Area: an Heteroscedasticity Autocorrelation Robust Assessment

Fabio Profumo*
University of York

11/10/2018

Abstract

In this paper, I perform real-time forecast evaluation of the European Survey of Professional Forecasters (ECB SPF) using the Diebold and Mariano test for equal forecast accuracy with different loss functions and competing point forecasts from different simple benchmark models. As macroeconomic historical data is subject to revision, I take it into account when constructing competing forecasts: in practice, rational forecasters use information available at the time they have to produce a forecast and the same should be done when estimating and forecasting a realistic comparison benchmark model. In addition, revision allows me to compare forecasts to different releases of historical data. As the sample size for ECB SPF is small and this affects size performances of the Diebold and Mariano test, I use fixed b and fixed m asymptotics that proved to alleviate size distortion in small samples. Results for the whole sample are not generally affected by revision of historical data and loss functions; ECB SPF does not seem to outperform benchmark models and, in some cases, benchmarks seem to perform better. Results in the sub sample before the 2008 financial crisis are similar to those for the full sample exercise while in the post crisis sample, rejection in favour of SPF forecasts is more frequent and results are more affected by the choice of loss function.

Keywords: Diebold and Mariano test, long run variance estimation, fixed-smoothing asymptotics, Heteroscedasticity Autocorrelation Robust (HAR) inference, SPF, Real-time forecast evaluation, Hypothesis testing

JEL Classification: C12, C32, C51, C53, E17

*Department of Economics and Related Studies, University of York, Heslington, YO10 5DD, UK. fabio.profumo@york.ac.uk. The support of the ESRC grant ES/J500215/1 is gratefully acknowledged.

1 Introduction

The objective of this paper is to assess whether simple and naïve benchmarks models can provide better forecasts than the European Central Bank's Survey of Professional Forecasters; for this purpose, I perform a fully real-time mean point forecast evaluation of ECB SPF using the Diebold and Mariano test for equal forecast accuracy with different loss functions. ECB SPF provide key insights about the euro area and they constitute an authoritative source about private sector expectations.

All forecasts, in general, are used by central banks, academic institutions, consumers and firms to make decisions and set future policy. ECB SPF is a quarterly survey about forecasts of fundamental economic variables: inflation, unemployment rate and real GDP growth for the euro area. Given the influence and the amount of resources involved in collecting and managing such surveys, it is essential to understand if it is solid and more effective than a simple forecast model. Doing so poses a series of decisions like the choice of loss function, benchmark models and vintage of the realisation for the target variable.

A possible way to evaluate forecasts is using informal graphics methods as in Theil (1958) which suggests using scatter plots of the forecast against the outcome to understand the magnitude of forecast errors. More formal approaches were proposed by Wilson (1934) which uses correlation between forecasts and realisations, Mincer and Zarnowitz (1969) that proposes a test for forecast unbiasdness and Fair and Shiller (1989, 1990) that examine the information content of ex-ante forecasts among others.

When two competing forecasts are available for the same variable of interest, Chong and Hendry (1986) propose a test for forecast encompassing while Diebold and Mariano (1995) suggest a test for equal forecast accuracy. In small samples, this test suffers from size distortion as noted by Clark (1999) and others. This issue can be alleviated using a heteroscedasticity autocovariance robust approach for the long run variance estimator required in the computation of the test statistic such as fixed b asymptotics by Kiefer and Vogelsang (2005) and fixed m asymptotics by Hualde and Iacone (2015) which proved to be good in delivering correctly sized tests in combination to a quadratic loss function as simulations in Coroneo and Iacone (2015) and Harvey, Leybourne and Whitehouse (2016) show. A challenge in forecasts evaluation comes from the fact that realisations of macroeconomic variables are revised often as discussed in Croushore and Stark (2001), Stark and Croushore (2002), Croushore (2006) and Clark and McCracken (2009). Revisions need to be taken into account when producing competing forecasts, for instance, using the same information set available to forecasters at the point in time when forecasts were made and also when selecting the value to consider as realised of the target variable. Moreover, the asymptotic distribution of forecast accuracy tests can change in the presence of predictable data revision. Also, robustness of evaluation to different loss functions is an important aspect of the analysis as the loss of agents is usually unknown. Results in Elliott, Komunjer and Timmermann (2008) indicate that this can potentially

cause troubles as forecast rationality tests are not robust to loss function specification and the same thing could potentially happen in tests for equal forecast accuracy although various reasons support the adoption of loss functions different from the usual ones, for instance, Capistrán and Timmermann (2009) provide several arguments supporting the choice of an asymmetric loss function such as asymmetries in costs arising from over and under predicting, psychological causes and strategic reasons.

Several works are available about evaluation of SPF: D'Agostino, Giannone and Surico (2006) use relative mean square errors for US SPF from 1975.Q1 to 1999.Q4 and find that predictive ability declined after the 80s. Boero, Smith and Wallis (2008) focus on Survey of External Forecasters from May 1996 to November 2005 and identify systematic overpredictions for inflation and GDP growth in the UK. Stark (2010) performs a real-time forecast evaluation of the US SPF with the Diebold and Mariano test over the sample 1985.Q1 -2007.Q4 finding a general good predictive ability which deteriorates as the forecast horizon gets longer. Coroneo and Iacone (2015), instead, perform an evaluation of both US SPF from 1985.Q1 to 2014.Q4 and EU SPF from 2006.Q1 to 2016.Q4 using the Diebold and Mariano test and fixed smoothing asymptotics to account for the small sample size of the EU SPF. Their findings confirm previous literature results. Also Demetrescu, Hanck and Kruse (2018) evaluate predictive accuracy of US SPF from 1969.Q1 to 2017.Q2 finding that after the Great Moderation the predictive accuracy has sensibly decreased. Bowles, Friz, Genre, Kenny, Meyler and Rautanen (2011) evaluation of EU SPF for real GDP growth and unemployment from 1999.Q1 to 2008.Q4 shows a moderate superiority of surveys over benchmarks however authors report that findings may be subject to small sample bias. To address the small sample of EU SPF issue, in this work, I perform real-time forecast evaluation of ECB SPF mean point forecasts with the Diebold and Mariano test for equal forecast accuracy, using fixed b and fixed m asymptotics to help obtain correctly sized test, with different loss functions and different vintages for the realised values of target variables (inflation, unemployment and real GDP Growth). Competing forecasts are produced keeping into account revision in historical data and using three different benchmarks: a simple Random Walk, an indirect autoregressive model and a direct autoregressive model. In the full sample exercise (2002.Q1 – 2010.Q3), forecasts are compared with actual values at different releases (vintages): first release, four releases after the first, twenty releases after the first and latest available release at 01/02/2018, while in pre-crisis (2002.Q1 – 2007.Q4) and post-crisis (2008.Q1 – 2012.Q4) samples I could only use the latest available release.

In general, results are not too affected by revision in historical data and loss function used. In terms of benchmark models, the Random Walk model seems to perform better than other models especially for long horizon forecasts.

The remainder of this paper is organised as follows. In section 2, I provide a description of the ECB SPF and of the real-time database, in section 3 I describe models used to generate competing forecasts, section 4 talks about loss functions involved in the eval-

uation process and section 5 gives a short outline of Diebold and Mariano test with fixed smoothing asymptotics. Section 6 describes the evaluation exercise and section 7 concludes.

2 Dataset

Mean point forecasts of Harmonised Index of Consumer Prices (HICP), Unemployment rate and real GDP growth are taken from the European Central Bank Survey of Professional Forecasters. The ECB SPF was started in 1999 with the aim to gather information about private sector expectations and assess the credibility of the policy of the new central bank founded the year before. It contains forecasts for three main economic indicators:

1. Inflation: defined as the year on year percentage change of the HICP published by Eurostat,

2. GDP: Real gross domestic product growth is defined as the year on year percentage change of real GDP, based on standardised ESA definition,

3. Unemployment: the unemployment rate refers to Eurostat's definition and it is calculated as percentage of the labour force.

The ECB's SPF is conducted four times per year, in the second half of the middle month of each quarter and, from the last quarter of 2001, in the second half of the first month of the quarter. A list of deadlines for reply to the survey is available on the ECB website. The ECB's SPF questionnaire is regularly submitted to a panel of forecasters (about 80 institutions with an average of 60 responses each round), all of the participants are experts affiliated with financial or non-financial institutions based within the EU and have been chosen to form an heterogeneous group in order to guarantee the representativeness and independence of the expectations collected. Panelists need to be experts in macroeconomics ad have previous forecasting experience for the euro area. The survey is about the Euro Area but respondents can be also based in the whole European Union, including countries which are not using the euro as currency. Table 1 shows timings, information available to forecasters and forecasts requested for each quarterly survey: for inflation and unemployment rate, forecasters are asked to forecast a specific month one year, two years and five years ahead from the latest available realisation of the target and not from the survey date. For real GDP growth, forecasts are always referred to h years ahead of the latest information available but these forecasts are about quarters and not about a specific month. Although information available to forecasters is the one reported in table 1, in the case of inflation, at the survey deadline 2007.Q1, the latest realisation available to forecasters was January 2007 instead of December 2006. In my exercise, I use the exact information forecasters had available at the survey deadline as my aim is to perform a fully real-time exercise. As displayed in figure 1a, revision for inflation is

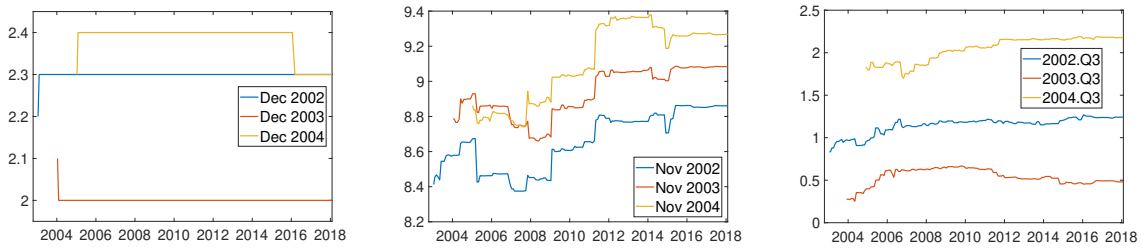
Inflation					
Survey	Month	Info available	Forecast 1 year	Forecast 2 years	Forecast 5 years
Q1.Y	1.Y	12.Y-1	12.Y	12.Y+1	12.Y+4
Q2.Y	4.Y	3.Y	3.Y+1	3.Y+2	3.Y+5
Q3.Y	7.Y	6.Y	6.Y+1	6.Y+2	6.Y+5
Q4.Y	10.Y	9.Y	9.Y+1	9.Y+2	9.Y+5

Unemployment					
Survey	Month	Info available	Forecast 1 year	Forecast 2 years	Forecast 5 years
Q1.Y	1.Y	11.Y-1	11.Y	11.Y+1	11.Y+4
Q2.Y	4.Y	2.Y	2.Y+1	2.Y+2	2.Y+5
Q3.Y	7.Y	5.Y	5.Y+1	5.Y+2	5.Y+5
Q4.Y	10.Y	8.Y	8.Y+1	8.Y+2	8.Y+5

Real GDP Growth					
Survey	Month	Info available	Forecast 1 year	Forecast 2 years	Forecast 5 years
Q1.Y	1.Y	Q3.Y-1	Q3.Y	Q3.Y+1	Q3.Y+4
Q2.Y	4.Y	Q4.Y-1	Q4.Y	Q4.Y+1	Q4.Y+4
Q3.Y	7.Y	Q1.Y	Q1.Y+1	Q1.Y+2	Q1.Y+5
Q4.Y	10.Y	Q2.Y	Q2.Y+1	Q2.Y+2	Q2.Y+5

Table 1: SPF Timing: the survey is produced quarterly but forecasts about inflation and unemployment are about a specific month: end of quarter month and middle of quarter monnth respectively. For real GDP growth, forecasts are about quarters. Forecasts horizons are 1 year, 2 years and 5 years ahead from the latest information available and not from the date of the survey. Y is the year considered.

negligible and it usually happens right after the first release.
In the case of unemployment, there is more revision and forecasters may not use newly available information because they are aware it is not reliable. In this case, I try both keeping and ignoring the additional information to obtain benchmark forecasts and results, available upon request, do not change. In this regard, surveys affected by this phenomenon are 2007.Q1 for which forecasters had one more realisation and 2004.Q2, 2008.Q4 and 2009.Q3 for which forecasters had one realisation less.
For real GDP growth, the latest information available is the one expected but it has already been revised once except in the case of survey 2002.Q2 and I always use the latest revision available at the survey deadline.
A special questionnaire was sent in September 2013 asking participants about their forecasting practices: responses indicate that forecasts are based on one or more model to cross check results but, especially for long term forecasts, judgment plays an important role with one third of respondents reporting that their forecasts are essentially judgement based. Moreover, the majority of participants reported the importance of judgement has increased following the financial crisis. For more information on surveys see Garcia (2003) and Bowles, Friz, Genre, Kenny, Meyler and Rautanen (2007).
A thorough analysis of responses is provided by Garcia and Manzanares (2007) in which a bias towards favourable predictions is discovered for all forecast horizons.



(a) Inflation (b) Unemployment (c) Real GDP Growth

Figure 1: Revision in historical data plotted from January 2003 to February 2018

In this paper, I consider inflation, unemployment rate and real GDP growth forecasts in surveys from 2002.Q1 until 2010.Q3 for a total of 35 observations obtained from the ECB website. Realizations are taken from the Real-time Database for the Euro Area available on the European Central Bank Statistical Data Warehouse; realised data about inflation spans from December 2002 to June 2015 (end of quarter months of the HICP annual growth), about unemployment, from November 2002 to May 2015 (middle of quarter months of the unemployment rate) and about real GDP growth, from 2002.Q3 to 2015.Q1.
Historical data is subject to revision, scheduled or not, caused by new data available, changes in definitions and classifications or correction of clerical mistakes. A real-time database is a collection of historical realisations and their revisions. For the United States,

the database was created by Croushore and Stark (2001) with data from November 1965 while, for the Euro Area, the real-time database was built by Giannone, Henry, Lalik and Modugno (2012) starting from January 2001.
Revisions should be taken into account when constructing benchmark forecasts to evaluate forecast accuracy. As in Stark (2010), when estimating and forecasting a benchmark model for comparison, I use the same data available to forecasters when they had to submit their forecast.

Mankiw and Shapiro (1986) argue that revision is most likely caused by unforecastable new information not known at time forecasts were made. Following Clark and McCracken (2009), which show this kind of revisions have usually no effect on asymptotic distribution of tests, I neglect its influence on asymptotics.

Figure 1 shows the effect of revision in realised data. For instance, the HICP annual growth rate for December 2002 was initially released on the 15/01/2003 at 2.2 and after revisions, it has been amended and kept to 2.3 until the 15/01/2018. For unemployment in November 2002, the first release was 8.41 on the 15/01/2003 and the last release available for the same month in my dataset is 8.86. For real GDP growth, the first release of the third quarter of 2002 was 0.83, this figure has been changed and my latest release is 1.24. The effect of revision is quite important in unemployment an Real GDP growth but minor in inflation confirming the findings of Giannone, Henry, Lalik and Modugno (2012). Revision patter is similar for US data: there is small or no revision for inflation, smaller revision than in Europe for Unemployment and bigger revision than in Europe for real GDP.
To account for a possible structural break caused by the financial crisis, I also repeat this exercise on two survey sub-samples: from 2002.Q1 to 2007.Q4 and from 2008.Q1 to 2012.Q4. Due to the shortage of revisions, I can only perform this evaluation using the current release of realised data but I can still consider revisions available at surveys deadlines to construct benchmark forecasts to retain the full real-time feature of this analysis.

3 Benchmarks

Following the approach by Stark (2010), I generate competing forecasts from three naive benchmark models taking into account the revision of historical data available at the time forecasters had to submit their forecasts each quarter in order to construct credible competing forecasts and perform a fair and consistent comparison. For every quarter, I check the deadline for replying to the survey round and I estimate benchmark models using the same information set forecasters had available before that date.

3.1 Random Walk

Let h be the forecast horizon, t be the date of the latest information available and V be the information set available at the deadline of the survey. The first benchmark model is a Random Walk

$$y_{t+h}^V = y_t^V + u_{t+h} \tag{1}$$

and the forecast h steps ahead is given by

$$\hat{y}_{t+h}^V = y_t^V \tag{2}$$

where y_t^V is the last historical realization available when the forecast is produced at the time of the survey. The forecast of this benchmark is the same no matter the forecast horizon.

As shown in Atkeson and Ohanian (2001) and Balcilar, Gupta, Majumdar and Miller (2015), the Random Walk model is hard to beat in forecasting inflation and real GDP although being trivial.

3.2 Indirect Autoregressive Model

The second benchmark model is a one-period univariate, indirect autoregressive model (IAR)

$$y_t^V = \theta_0 + \sum_{j=1}^{P(V)} \theta_j y_{t-j}^V + u_t \tag{3}$$

For each survey, parameters are estimated using the last 30 quarterly observations available at vintage V. The lag length $P(V)$ is chosen using the Bayesian Information Criterion re-estimated each survey round, the maximum lag is 4.

Forecasts are obtained recursively according to

$$\hat{y}_{t+h}^V = \hat{\theta}_0 + \sum_{j=1}^{P(V)} \hat{\theta}_j \hat{y}_{t-j+h}^V \tag{4}$$

Marcellino, Stock and Watson (2006) suggest that IAR works best when the model is correctly specified.

3.3 Direct Autoregressive Model

To account for model misspecification, I also use a Direct Autoregressive model (DAR) as a third benchmark

$$y_t^V = \theta_0^{(h)} + \sum_{j=1}^{P(h,V)} \theta_j^{(h)} y_{t-j-h}^V + u_t^{(h)} \tag{5}$$

which tends to be more robust to model misspecification according to Schorfheide (2005) and Bhansali (2002) among others.
For each survey, parameters are estimated using the last $30 - h$ quarterly observations available at vintage V. The lag length $P(h, V)$ is chosen using the Bayesian Information Criterion and re-estimated at each survey round and every forecast horizon, the maximum lag is 4.
Forecasts are obtained directly from

$$\hat{y}^V_{t+h} = \hat{\theta}^{(h)}_0 + \sum_{j=1}^{P(h,V)} \hat{\theta}^{(h)}_j y^V_{t-j} \tag{6}$$

4 Loss Functions

Forecast evaluation with the Diebold and Mariano test involves the use of a loss function of forecast errors $e^V_{t+h} = \hat{y}_{t+h} - y^V_{t+h}$ where y^V_{t+h} is the realisation of the target variable at vintage V at time $t+h$ and $\hat{y}_{t+h}$ is its forecast for time $t+h$; most common loss functions are the Quadratic loss, defined as

$$L(e^V_{t+h}) = {e^V_{t+h}}^2 \tag{7}$$

and the Absolute loss

$$L(e^V_{t+h}) = |e^V_{t+h}|. \tag{8}$$

Both these functions are commonly used in the literature as they are well known and easy to deal with, they are both symmetric, bowl shaped, differentiable everywhere (except in zero for the Absolute loss) and unbounded from above. They also satisfy all three Granger (1999) properties: minimal loss of zero, loss always positive or equal to zero, non increasing for negative forecast errors and non decreasing for positive forecast errors. Large forecast errors are highly penalised but while for the Quadratic loss penalty increases quadratically, for the Absolute loss, it increases linearly.
In addition to the two naive Quadratic and Absolute loss functions, I use the Linex function by Varian (1975) which is asymmetric but it is still differentiable everywhere and it takes the form

$$L(e^V_{t+h}) = \exp(\alpha e^V_{t+h}) - \alpha e^V_{t+h} - 1 \tag{9}$$

where α is a scalar that controls the aversion towards positive ($\alpha > 0$) or negative forecast errors ($\alpha < 0$). The choice of this parameter has to be done according to costs arising from overpredicting or underpredicting the target variables.

With this type of loss function, it is possible to weight forecast errors according to their sign; I compare forecasts using both α positive and negative as α is set at values 1 and

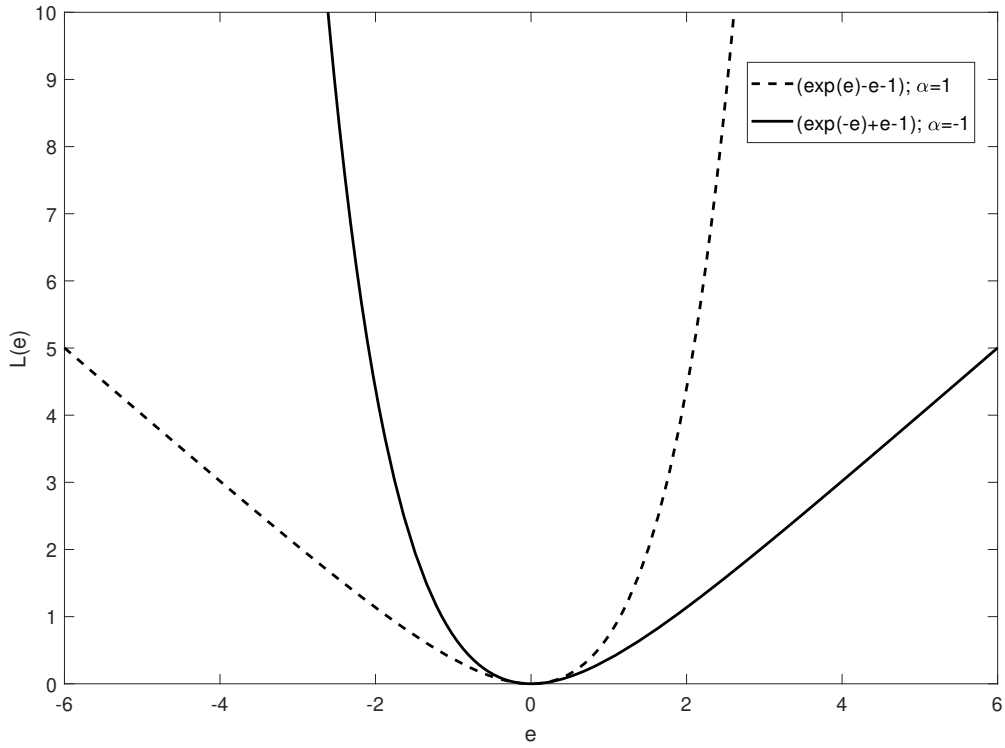


Figure 2: Linex loss function $L(e^V_{t+h}) = \exp(\alpha e^V_{t+h}) - \alpha e^V_{t+h} - 1$ with $\alpha = -1$ and $\alpha = 1$ where the forecast error is defined as $e^V_{t+h} = \hat{y}_{t+h} - y^V_{t+h}$ with y^V_t the realisation of the target variable at vintage V and $\hat{y}_t$ its forecast.

-1. Positive forecast errors occurs when the realised value for the target variable is bigger than its corresponding forecast and vice-versa for negative forecast errors.

5 DM Test and Fixed-smoothing Asymptotics

To evaluate the predictive performance of EU SPF, I use the DM test for the null of equal forecast accuracy by Diebold and Mariano (1995).
Given forecast errors defined as $e^V_{t+h} = \hat{y}_{t+h} - y^V_{t+h}$, $\hat{y}_{t+h}$ the forecast h steps ahead of the actual y_{t+h}, the loss differential is $d_{t+h}(L)^V = L(e^{V,B}_{t+h}) - L(e^{V,SPF}_{t+h})$ with $L(e^{V,B}_{t+h})$ the loss function evaluated at forecast errors from benchmark models and $L(e^{V,SPF}_{t+h})$ the loss function evaluated at forecast errors from SPF.
The unfeasible test statistic is

$$\sqrt{T}\frac{\bar{d}^V - \mu^V}{\sigma} \xrightarrow{d} N(0,1) \tag{10}$$

where $\bar{d}^V = \frac{1}{T}\sum_{t=1}^{T} d^V_t(L)$ and $\mu^V = E(d^V_t)$. The null hypothesis is $H_0 : \mu = 0$.
If the estimator for the long run variance σ^2 is consistent, the limiting distribution is

still a standard Normal. The estimator suggested by Diebold and Mariano (1995) is a Weighted Covariance Estimator

$$\hat{\sigma}^2_{WCE-DM} = \hat{\gamma}_0 + 2\sum_{j=1}^{h-1}\hat{\gamma}_j \tag{11}$$

with $\hat{\gamma}_j = \frac{1}{T}\sum_{t=1}^{T-j}(d_t^V(L) - \bar{d}^V)(d_{t+j}^V(L) - \bar{d}^V)$.
It is consistent as it is based on the assumption that $d_t^V(L) - \mu^V$ is a MA$(h-1)$, with h the forecast horizon. However it may generate negative estimates because of the rectangular kernel employed and this is troublesome.
Simulations in Diebold and Mariano (1995) also show large size distortion in small samples, it is possible to replace the kernel function, for instance using a Bartlett kernel, to get

$$\hat{\sigma}^2_{WCE-B} = \hat{\gamma}_0 + 2\sum_{j=1}^{T-1}k_{BART}(j/M)\hat{\gamma}_j \tag{12}$$

but this chamge does not seem to eliminate the size distortion issue as shown in Clark (1999).

Kiefer and Vogelsang (2005) suggest fixed b asymptotics as an alternative to small b asymptotics. Their assumption is based on taking $b = \frac{M}{T} \in (0, 1]$ fixed as $T \to \infty$ where M is the bandwidth. Under this assumption, $\hat{\sigma}^2$ is not consistent and not asymptotically unbiased, as a result, DM test statistic has a non standard distribution which depends on both b and the kernel choice.

$$\sqrt{T}\frac{\bar{d}^V - \mu^V}{\hat{\sigma}_{WCE-B}} \Longrightarrow \Phi_{BART}(b) \tag{13}$$

$\Phi_{BART}(b)$ is characterised in Kiefer and Vogelsang (2005) and a cubic equation is provided for critical values.

An alternative set of estimators for the long run variance is available when moving from time domain to frequency domain and using periodograms instead of autocovariances.
In general, a Weighted Periodogram Estimator for σ^2 is

$$\tilde{\sigma}^2 = 2\pi\sum_{\tau=1}^{T/2}K_M(\lambda_\tau)I(\lambda_\tau) \tag{14}$$

where $K_M(\lambda_\tau)$ is a symmetric kernel function, $I(\lambda_\tau) := |\frac{1}{\sqrt{2\pi T}}\sum_{t=1}^{T} d_t(L)e^{i\lambda_\tau t}|^2$ is the periodogram of $d_t(L)$ evaluated at $\lambda_\tau = 2\pi\tau/T$ Fourier frequencies for $\tau = 0, \pm 1, \ldots, \pm T/2$.
When a Daniell kernell is used, the long run variance estimator is

$$\hat{\sigma}^2_{WPE-D} = 2\pi\frac{1}{m}\sum_{\tau=1}^{m}I(\lambda_\tau) \tag{15}$$

where m is a function of M. When $m \to \infty$ as $T \to \infty$, the estimator is consistent and the limiting distribution of the DM test is still a standard Normal. This approach is referred to as large m asymptotics.
Hualde and Iacone (2015) suggest fixed m asymptotics; in this case, $\hat{\sigma}^2_{WPE-D}$ is no longer consistent but still asymptotically unbiased and

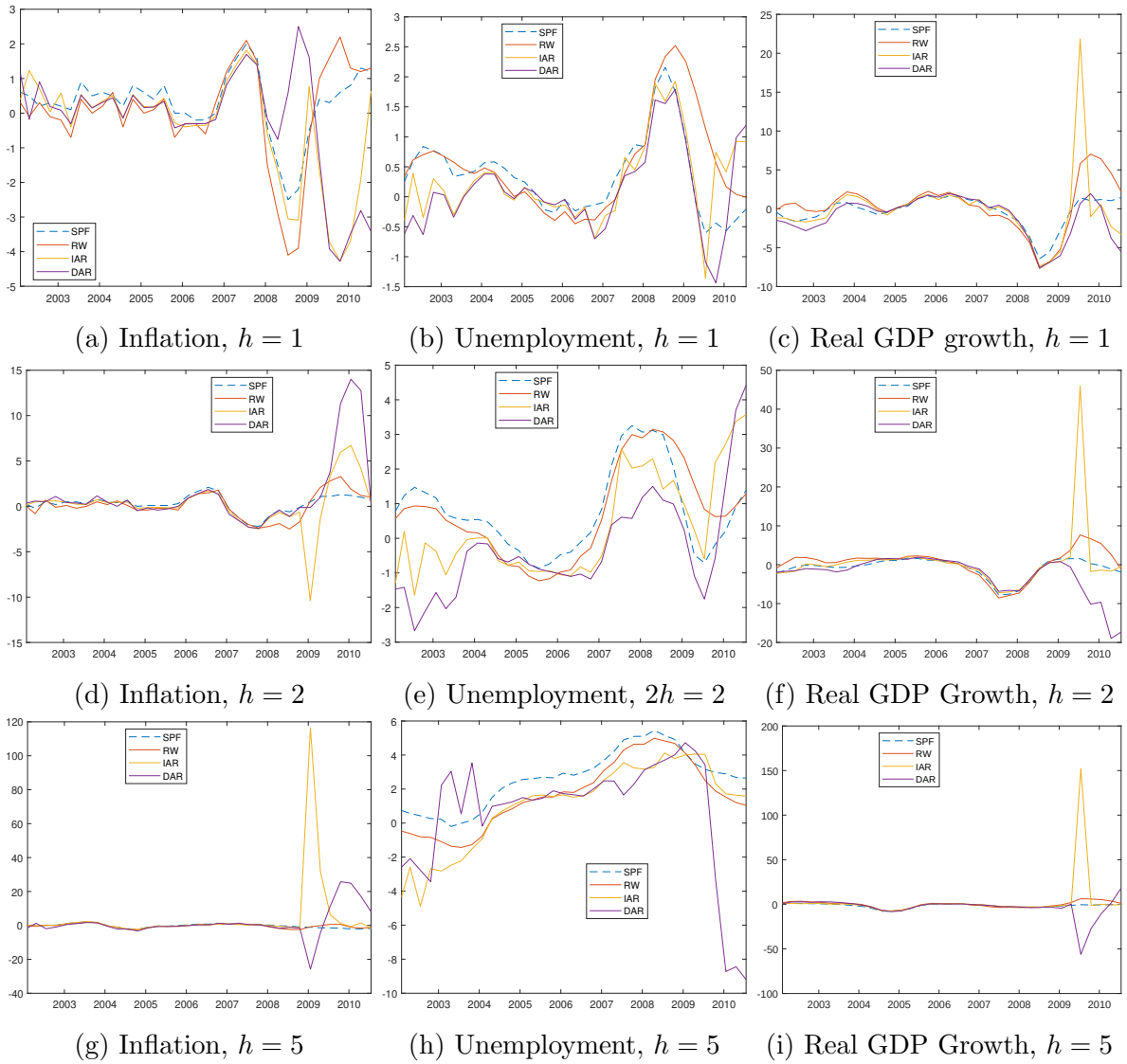
$$\sqrt{T}\frac{\bar{d}^V - \mu^V}{\hat{\sigma}_{WPE-D}} \xrightarrow{d} t_{2m} \tag{16}$$

without the need for a non standard distribution to be simulated.
Both fixed b and fixed m approaches bring a remarkable improvement in size but with a size-power trade off as shown in Coroneo and Iacone (2015) and Harvey et al. (2016) with a Quadratic loss function.
I perform additional simulations with an Absolute loss function and a Linex loss function and these exercises show the same improvement in size.[1]

6 Evaluation of SPF

To test the null hypothesis of equal forecast accuracy on European SPF, I perform the Diebold and Mariano test using a Weighted Covariance long run variance estimator with Bartlett kernel (WCE bandwidth $M = T^{1/2}$) and a Weighted Periodogram long run variance estimator with Daniell kernel (WPE bandwidth $m = T^{1/3}$). Bandwidths are chosen as advised in Coroneo and Iacone (2015) and critical values are from fixed b asymptotics by Kiefer and Vogelsang (2005) and fixed m asymptotics by Hualde and Iacone (2015) respectevely which have better size performances in small sample as shown in Coroneo and Iacone (2015) and Harvey, Leybourne and Whitehouse (2016).
I test one year, two years and five years ahead SPF survey average forecasts from 2002.Q1 to 2010.Q3 of the target variables against forecasts from a Random Walk model, IAR model and DAR model constructed taking into account revision of realised data. I repeat the test using different loss functions (quadratic, absolute and Linex losses) and alternative values of historical realisations (first release, four releases after the first, twenty releases after the first and latest release available at 01/02/2018).
Forecast errors for the last release of target variables are displayed in figure 3. For all the variables considered, it seems IAR and DAR models struggle to provide reliable forecasts from 2008 to 2010 while SPF and the Random Walk still fare quite good. This behaviour could suggest a structural break given by the financial crisis in 2008. For this reason, I also perform the exercises on two separate sub-samples, 2002.Q1 - 2007.Q4 and 2008.Q1 - 2012.Q4, with 24 an 20 observations respectively, only using the current release of actual realised data as it is the only one available for latest surveys.

[1] Results available upon request.



(a) Inflation, $h = 1$ (b) Unemployment, $h = 1$ (c) Real GDP growth, $h = 1$

(d) Inflation, $h = 2$ (e) Unemployment, $2h = 2$ (f) Real GDP Growth, $h = 2$

(g) Inflation, $h = 5$ (h) Unemployment, $h = 5$ (i) Real GDP Growth, $h = 5$

Figure 3: Forecast errors for current release (01/02/2018). h is the forecast horizon in years.

Performing the test on the full sample 2002.Q1 to 2010.Q3, I find that there is little effect of revision on inflation and real GDP growth and small, but more than in the US, on unemployment. Results of Diebold and Mariano test for the null of equal forecast accuracy at 5%, 10% and 20% significance levels in the case of inflation are displayed in figures 4 - 7. I cannot reject the null of equal forecast accuracy for all the benchmarks no matter the revision of historical data and the loss function I consider.

For five years ahead forecasts, the Random Walk always performs better than the other benchmark models as the test statistic is moving towards zero and when using a Linex loss with $\alpha = 1$ the test statistic becomes negative, still, the null of equal forecast accuracy cannot be rejected; for one and two years ahead forecasts, instead, the Random Walk performs slightly worst than other models. The null cannot be rejected even considering a significance level of 10% and only using a generous level of significance of 20%

I can reject the null hypothesis only on occasional instances in which benchmark models perform worst.
Turning to unemployment, figures 8 - 11 show test results; I cannot reject the null hypothesis at 5% significance level except for the case of a quadratic loss function being used on five years ahead forecasts of the Random Walk model. Also in this case, the Random Walk model seems to perform better than other benchmarks.
Revision has no effect on the outcome of the 5% test but as data is revised, the test statistic shifts downwards; this movement is less detectable when a Linex loss with $\alpha = 1$ is used.
Considering a level of significance of 10%, the effect of revision is more noticeable: for instance, using an absolute loss function to evaluate one year ahead forecasts, the null hypothesis is rejected for an IAR using the fourth release after the first and not for the others. However, the WPE-D case does not support this outcome.
In the case of real GDP growth, results are reported in figures 12 - 15. The DAR model seems to perform quite bad for one year ahead forecasts taking the test statistic to the rejection region when using the absolute loss function and Linex loss function with $\alpha = 1$; its performances improve for longer horizons. Using a Linex loss with $\alpha = -1$ makes the test statistic about the IAR model going negative and vice versa for DAR model in two years ahead forecasts.
Also, the Random Walk is not performing very well for two years ahead forecasts when combined with an absolute loss function.
Revision has little effect on test statistics and the only case in which it leads to rejection is the case of WCE-B, two years ahead forecasts evaluated using an Absolute loss function. In general, the IAR model seems to be the one that performs better but it still does not lead to rejection of the null in favour of the benchmark.
Considering larger significance levels, rejection occurs only for one year ahead forecasts and a Quadratic loss function and a Linex loss function with $\alpha = -1$.

On the whole, given the test results from the full sample, there is no strong evidence the ECB SPF outperforms benchmark models at all forecast horizons. In fact, Random Walk and the IAR model sometimes seem to predict better the target variable than the SPF. Especially the Random Walk takes the test statistic for unemployment with quadratic loss function to the rejection region for five years ahead forecasts.
Revision in historical realisations is also rather small to affect test statistics. A light effect of revision is visible for unemployment as the test statistic shifts down as the release updates but, in general, it does not change the outcome of evaluation. Considering real GDP growth two years ahead forecasts with an absolute loss function, revision takes the Random Walk test statistic in the 5% rejection region for the current release only but this result is not supported by the WPE case. Effect of revision is very similar to the one observed in the US.
Different benchmark models usually give the same results with the only exception given

by the DAR model for real GDP growth used in combination with an asymmetric loss function which takes the test to the rejection region. In general, the Random Walk performs quite well and slightly better for longer horizons which confirms the fact that the Random Walk model is hard to beat for predictions.
Differently from forecast rationality tests, there is no large difference of outcome across loss functions indicating that ignoring the actual loss function used by respondents and other agents involved while evaluating forecasts has little effect on the results of the test.

Turning to forecast evaluation on the first sample from 2002.Q1 to 2007.Q4, results are reported in figures 16, 18 and 20. The only case of strong rejection of the null hypothesis of equal forecast accuracy occurs in real GDP growth. Especially for two years ahead forecasts, the Random Walk has pretty bad performances and it takes the test statistic in the 5% rejection region. Five years ahead instead, the null can be rejected at 10% significance level in favour of the IAR model for WCE-B and at 20% significance level for WPE-D. For inflation there is no rejection except in five years ahead SPF forecasts against DAR forecasts evaluated with a Linex loss function with $\alpha = -1$ considering a significance level of 20%. Looking at unemployment, there is no rejection of the null for two years ahead forecasts; for one year ahead forecasts test statistics for IAR and DAR are close to the negative 20% rejection region and rejection is obtained using a Linex loss function with $\alpha = 1$. For five years ahead forecasts, the choice of loss function affects the outcome of the test: in the case of Random Walk forecasts, rejection is obtained in the negative region for a squared loss and a Linex loss with $\alpha = 1$.
Observing figures 17, 19 and 21 about the second sample from 2008.Q1 to 2012.Q4, rejection occurs more often than in the first sample and usually on the positive rejection region indicating that benchmarks performed worst than SPF. In particular, Random Walk test statistics for two years inflation forecasts are blatantly in the positive rejection region except for the Linex loss with $\alpha = 1$. IAR two years ahead test statistics are in the rejection region only for squared and absolute loss functions. For one year ahead forecasts, only Random Walk performs worst than SPF with squared, absolute and Linex with $\alpha = 1$ loss functions considering a 20% significance level; for the same significance level, five years ahead DAR forecasts evaluated with squared and absolute loss functions, appear worst than SPF forecasts. Similar results, although less evident, are obtained for real GDP growth one year and five years ahead forecasts in which both Random Walk and DAR forecasts perform worst than SPF for squared and absolute loss functions. For two years and five years ahead forecasts, rejection is only obtained considering a 20% significance level and symmetric loss functions; rejection is also obtained for five years ahead Random Walk forecasts evaluated with a Linex loss function with $\alpha = -1$. For unemployment, instead, evident rejection is for five years ahead forecasts but in the negative region showing that Random Walk and IAR forecasts had smaller forecasts errors than SPF. Results are more dependant on the choice of loss function: for IAR

forecasts, there is no rejection using a Linex loss with $\alpha = 1$ while there is rejection with other functions; for Random Walk forecasts, there is no rejection with absolute loss and Linex loss with $\alpha = -1$. For two years ahead, IAR and DAR forecasts take the test statistic around the 20% positive rejection region while Random Walk forecasts have on average the same forecast errors as the SPF. One year ahead, Random Walk is on average equivalent to SPF while IAR and DAR perform generally worst than SPF leading to rejection.

Comparing results from 2002.Q1 to 2007.Q4, in which there is no strong rejection of the null hypothesis confirming findings of the full sample while, to the ones from the second sample from 2008.Q1 to 2012.Q4, it appears rejection occurs more often and in favour of SPF forecasts. This can be the results of respondents having improved their ability to forecast with time and from the fact that, after the financial crisis, forecasts are not produced from internal models alone but also complemented by judgement of forecasters which seems to add value. In addition, the outcome of the test in the post crisis sample seems to be influenced by the choice of the loss function differently from the full sample and first sub sample.

Stark (2010) evaluates Survey of Professional Forecasters published by the Federal Reserve Bank of Philadelphia about the US from 1985.Q1 to 2007.Q4 using Root Mean Square errors and the Diebold and Mariano test. He finds that SPF are good forecasts and they always outperform all benchmark models. Revision has no effect on unemployment and very small on inflation while it has a strong effect on real GDP. In this case, SPF become more inaccurate as new revisions are released. Results are the same for all different benchmark models.

Comparing my findings to results from Stark (2010), US SPF appears to provide more accurate forecasts than ECB SPF but it should be noted that US SPF gives forecasts from the current quarter to the fourth quarter ahead (one year) while ECB SPF delivers forecasts from one year ahead (four quarters) to five years ahead (twenty quarters), so the forecast horizon is longer for European Surveys. Keeping these different horizons in mind, US SPF starts to deteriorate and loses advantage over benchmark models from the third/fourth quarter ahead which is the time I start to evaluate European SPF and I cannot reject the null of equal forecast accuracy.

In addition, Stark (2010) considers a different and longer period of time: his sample starts in 1985 and ends in 2007. There is evidence in the literature about SPF performing quite well in the past but loosing most of their predictive ability in the last two decades. In particular, D'Agostino, Giannone and Surico (2006), Coroneo and Iacone (2015) and Demetrescu, Hanck and Kruse (2018) find that after 1985, US SPF do not outperform a simple benchmark model and the predictive ability of surveys worsens as the time horizon gets longer. Considering that European surveys shortest horizon is one year ahead and the longest is five years ahead, my findings are consistent with the literature.

Bowles et al. (2011) find EU SPF about real GDP growth and unemployment to some

extent superior to basic benchmarks but their results are based on a very limited sample and they do not use any method to address the small sample bias like I do.

7 Conclusions

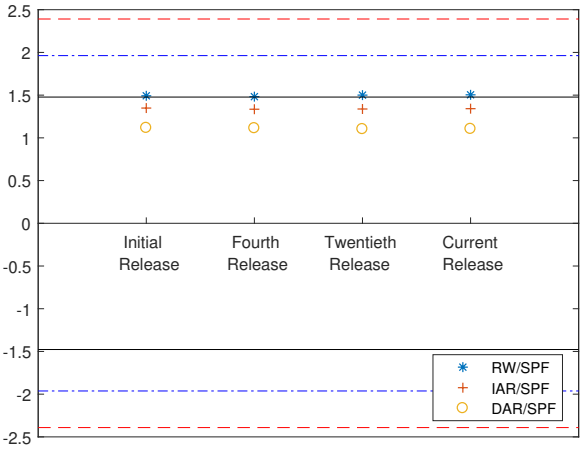
In this paper, I perform a fully real-time evaluation of EU SPF forecasts about inflation, unemployment rate and real GDP growth using the Diebold and Mariano test for equal forecast accuracy. Benchmark forecasts are taken from three simple models: Random Walk, Indirect Autoregressive and Direct Autoregressive. I consider different releases for the realisations of the target variables: first release, four releases after the first, twenty releases after the first and the latest release available. The sample available is small so, to account for small sample bias of this type of test, I use fixed b and fixed m asymptotics which have good size performance even in small samples.

Results for full sample show EU SPF do not outperform benchmark models. Similar results are obtained for the the sample before the 2008 financial crisis while in the post crisis sample, EU SPF recover predictive ability.

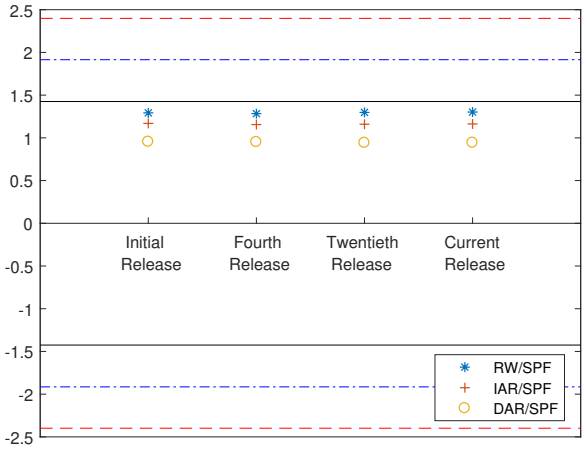
Many authors, such as Garcia and Manzanares (2007), Clements (2009), Engelberg, Manski and Williams (2009) and Clements (2010) among others, notice the presence of inconsistencies between the point estimates reported by forecasters and moments of their density forecasts, these inconsistencies have impacts on the average of point forecasts which I use in this work. In this light, the same evaluation study could be conducted on density forecasts which incorporate more information and do not suffer from bias.

Moreover, in this work, I do not consider parameter estimation error and, as a result, forecasting models strongly affected by this phenomenon could be judged inferior relative to models that are less strongly affected by parameter estimation error; in future work, real-time forecast evaluation can be performed with tests which retain the effect of estimation errors like the Giacomini and White (2006) test.

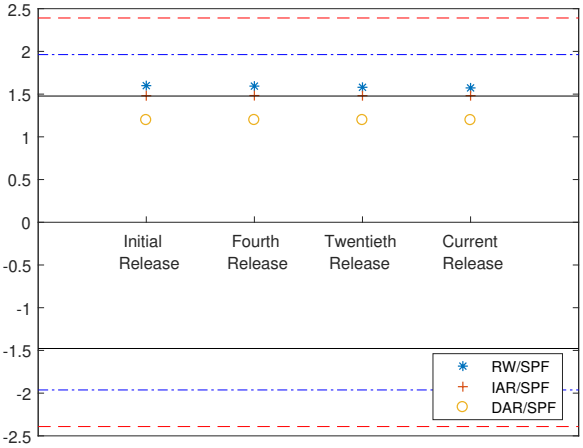
In addition, forecasts from benchmark models could be enhanced using data from multiple vintages as suggested in Clements and Galvão (2012) and others.



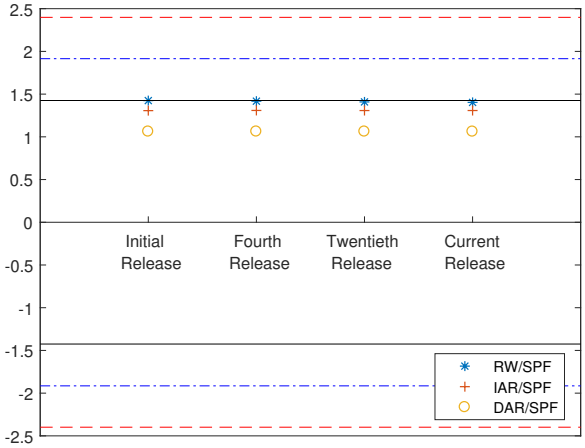
(a) WCE-B, 1 year ahead forecasts



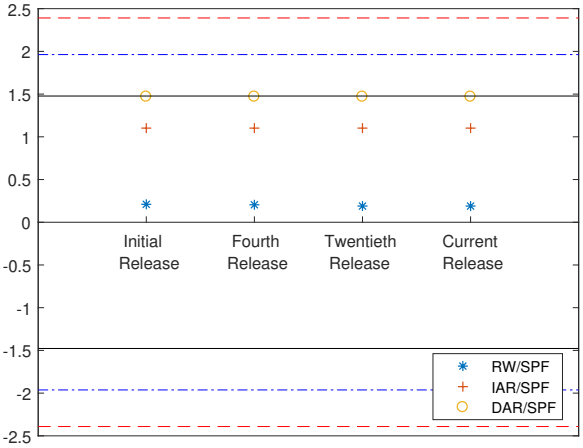
(b) WPE-D, 1 year ahead forecasts



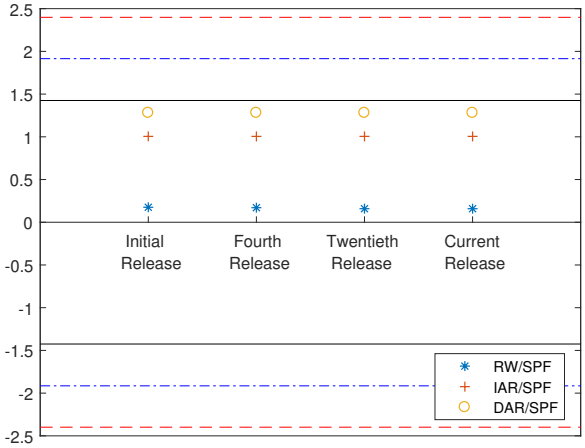
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

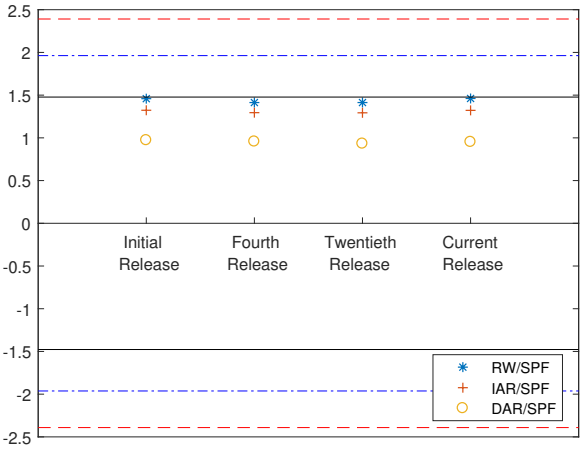


(e) WCE-B, 5 years ahead forecasts

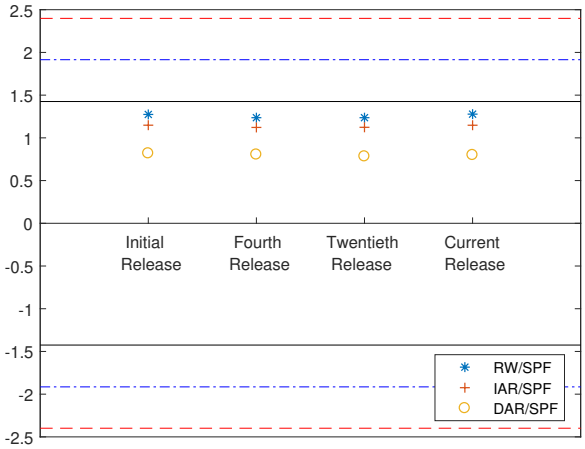


(f) WPE-D, 5 years ahead forecasts

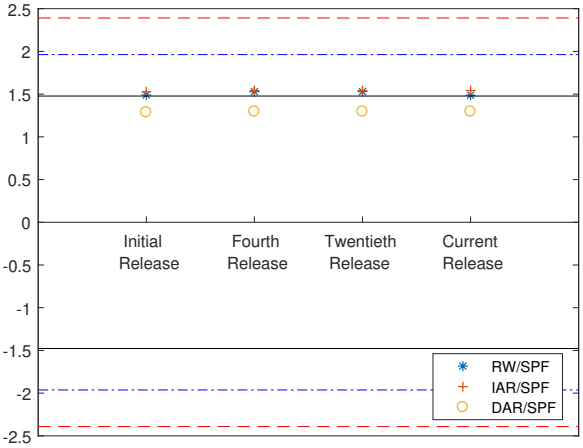
Figure 4: DM test statistic for inflation and quadratic loss function. Sample 2002.Q1 - 2010.Q3. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.3911; blue dash-dotted: 10%, 1.9626; black solid: 20%, 1.4774) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.3986; blue dash-dotted: 10%, 1.9147; black solid: 20%, 1.4253).



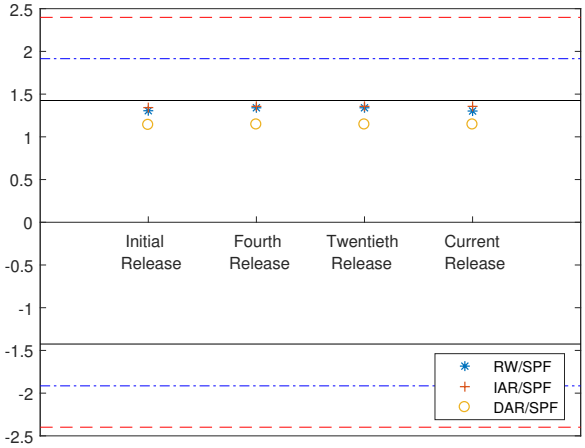
(a) WCE-B, 1 year ahead forecasts



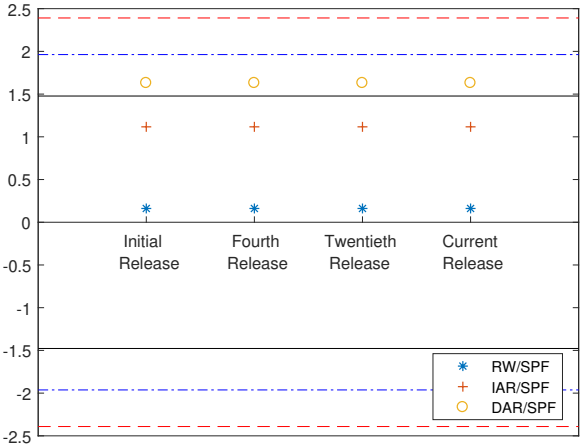
(b) WPE-D, 1 year ahead forecasts



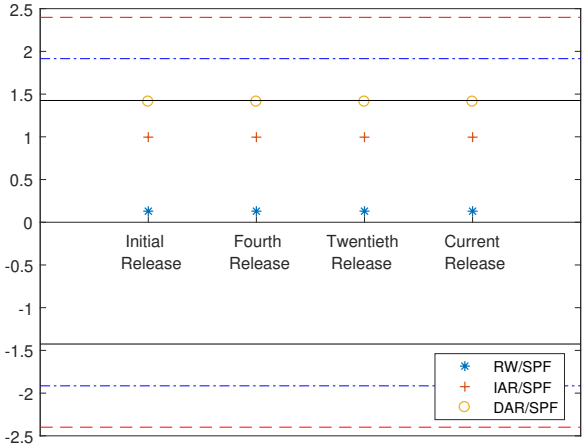
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

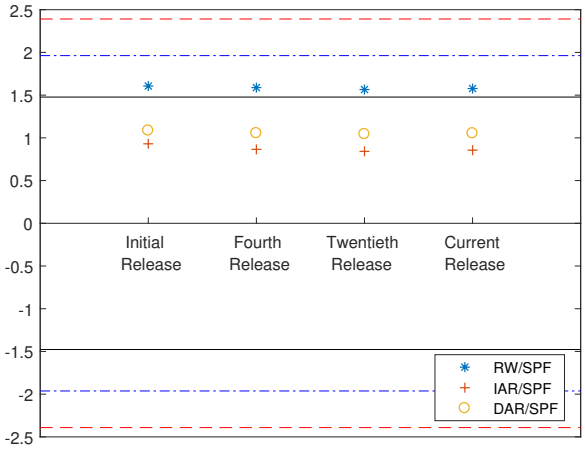


(e) WCE-B, 5 years ahead forecasts

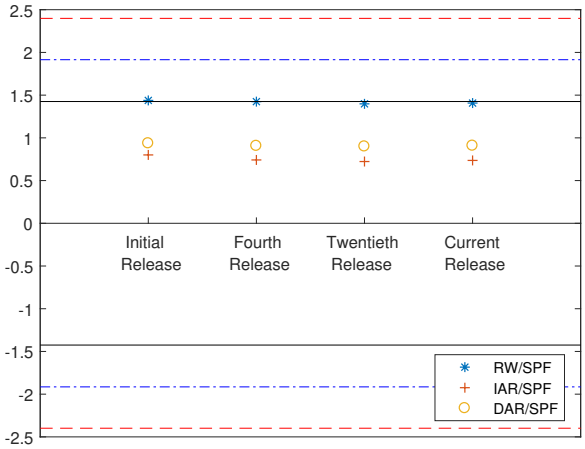


(f) WPE-D, 5 years ahead forecasts

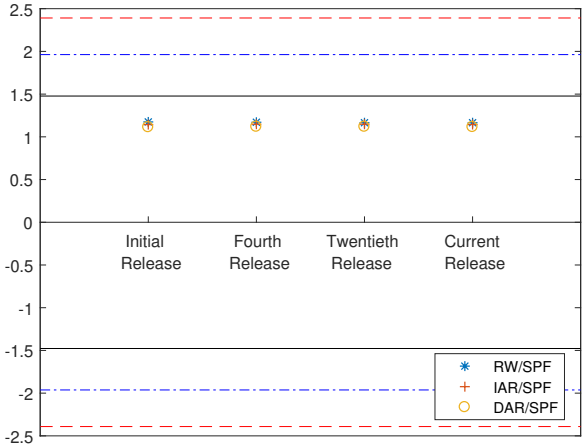
Figure 5: DM test statistic for inflation and absolute loss function. Sample 2002.Q1 - 2010.Q3. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.3911; blue dash-dotted: 10%, 1.9626; black solid: 20%, 1.4774) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.3986; blue dash-dotted: 10%, 1.9147; black solid: 20%, 1.4253).



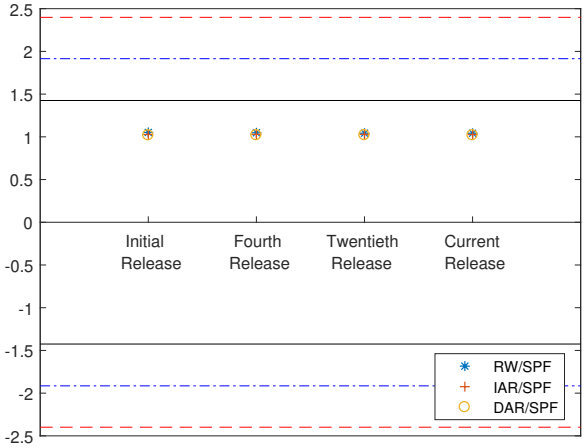
(a) WCE-B, 1 year ahead forecasts



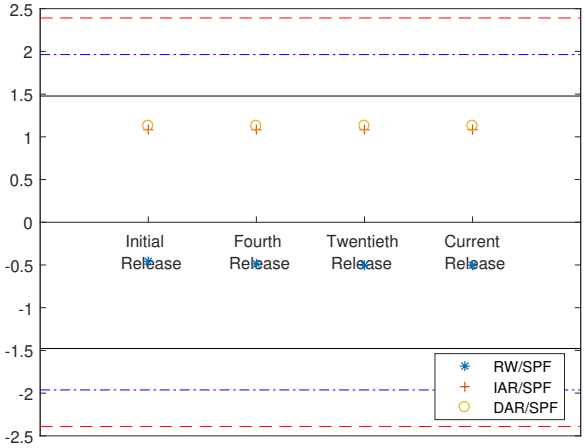
(b) WPE-D, 1 year ahead forecasts



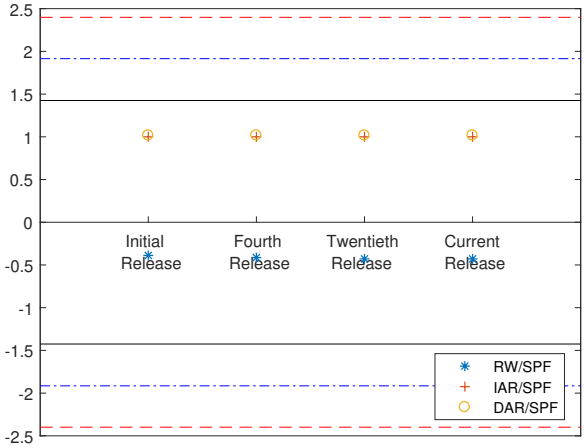
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

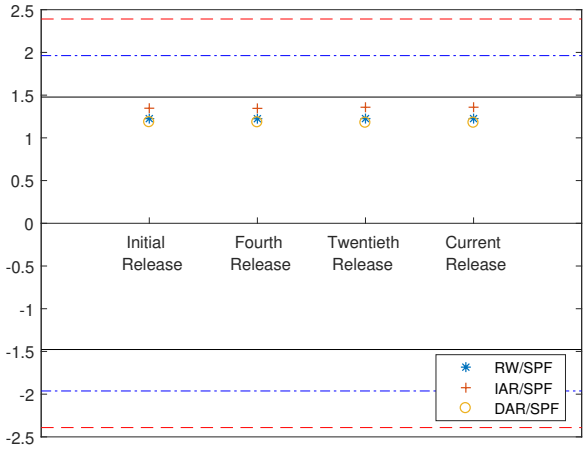


(e) WCE-B, 5 years ahead forecasts

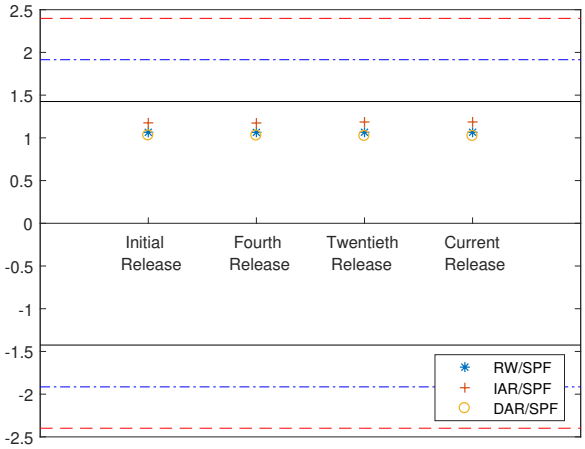


(f) WPE-D, 5 years ahead forecasts

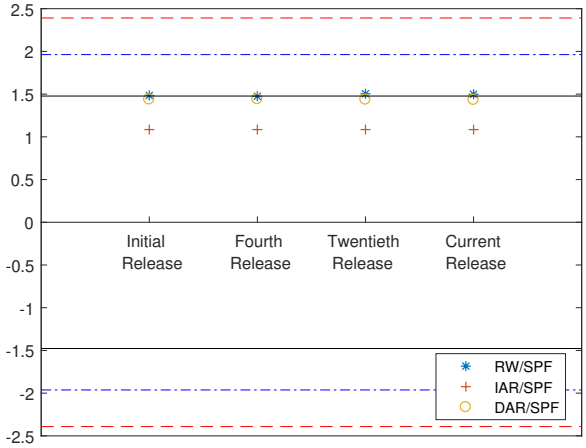
Figure 6: DM test statistic for inflation and Linex loss function $\alpha = 1$. Sample 2002.Q1 - 2010.Q3. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.3911; blue dash-dotted: 10%, 1.9626; black solid: 20%, 1.4774) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.3986; blue dash-dotted: 10%, 1.9147; black solid: 20%, 1.4253).



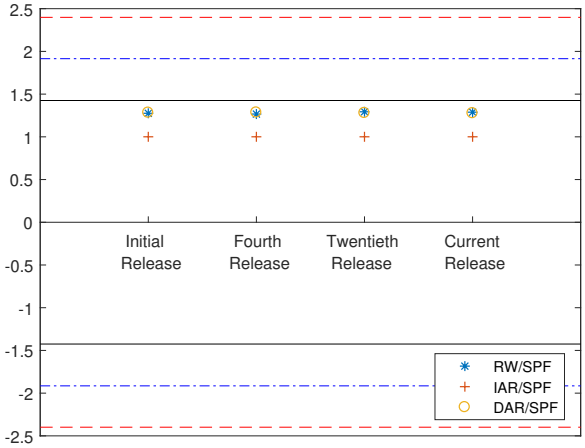
(a) WCE-B, 1 year ahead forecasts



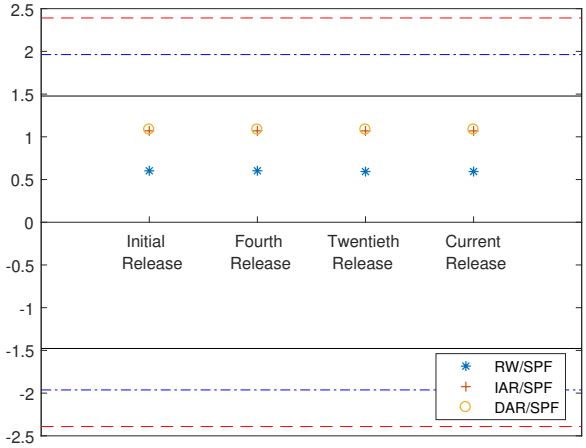
(b) WPE-D, 1 year ahead forecasts



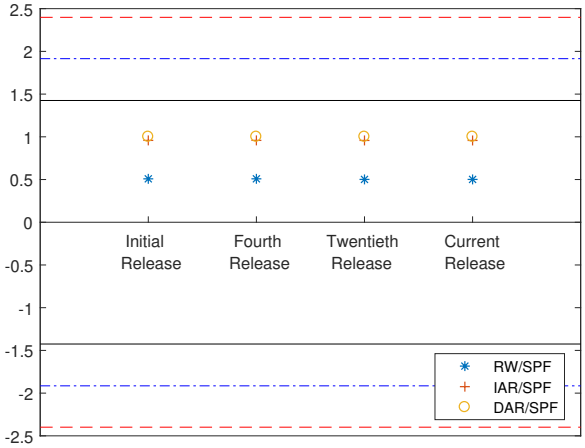
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

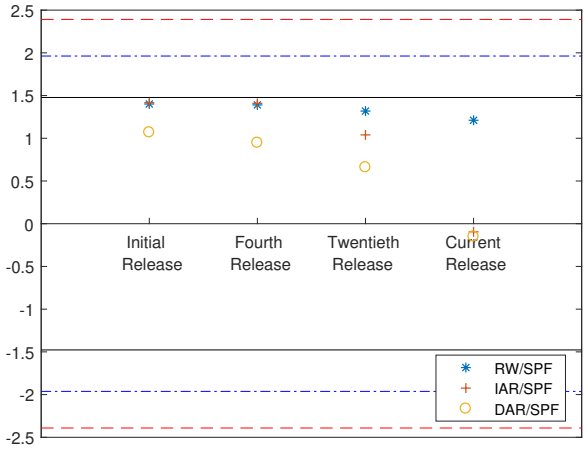


(e) WCE-B, 5 years ahead forecasts

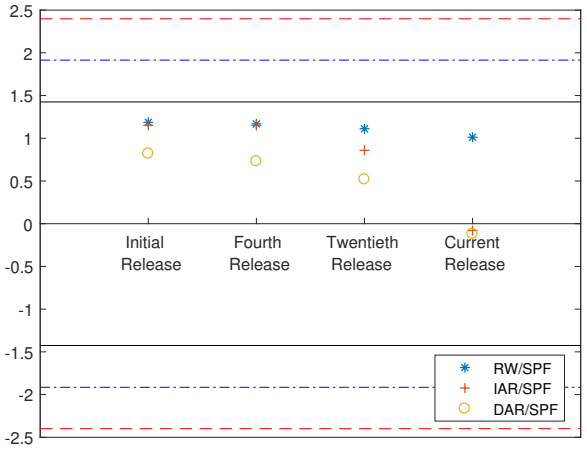


(f) WPE-D, 5 years ahead forecasts

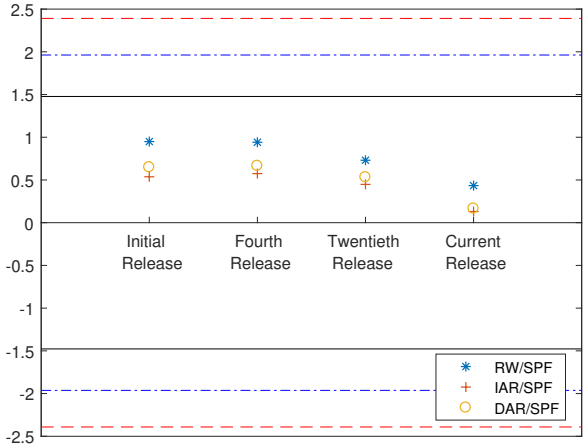
Figure 7: DM test statistic for inflation and Linex loss function $\alpha = -1$. Sample 2002.Q1 - 2010.Q3. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.3911; blue dash-dotted: 10%, 1.9626; black solid: 20%, 1.4774) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.3986; blue dash-dotted: 10%, 1.9147; black solid: 20%, 1.4253).



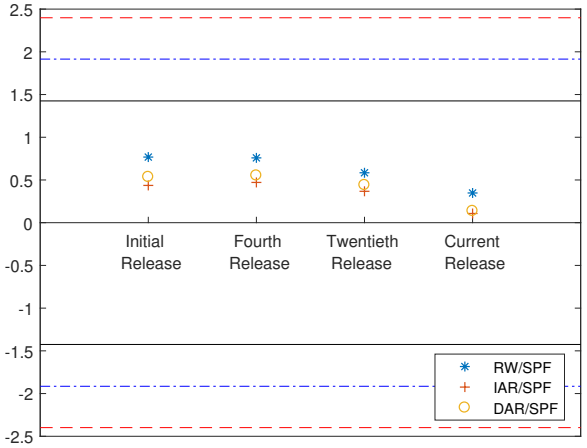
(a) WCE-B, 1 year ahead forecasts



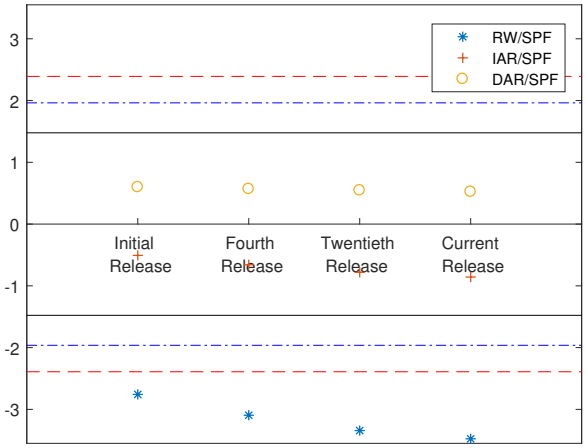
(b) WPE-D, 1 year ahead forecasts



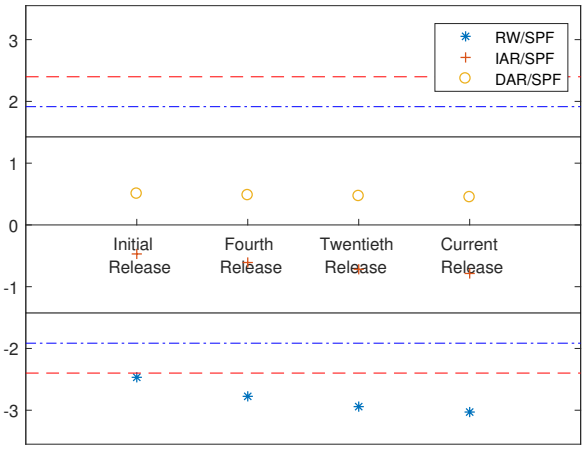
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

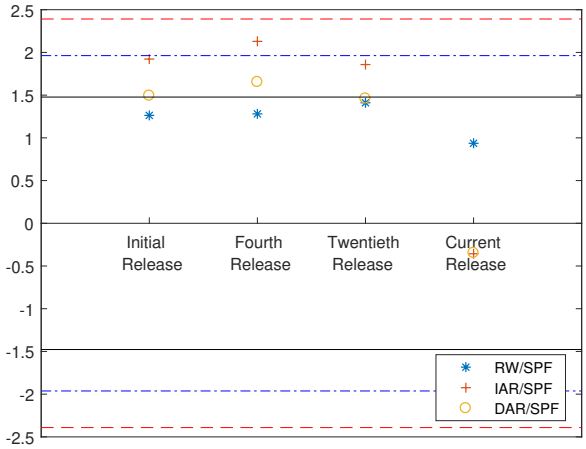


(e) WCE-B, 5 years ahead forecasts

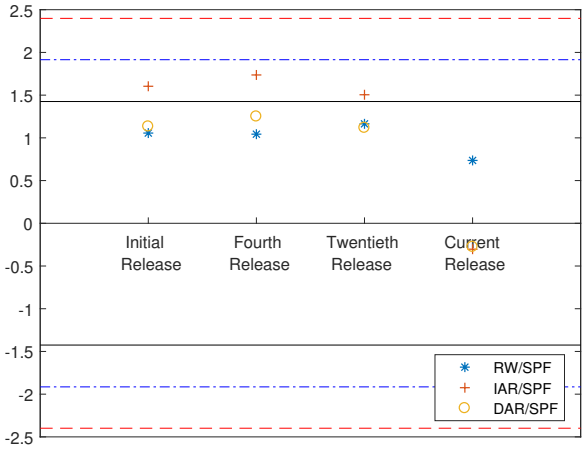


(f) WPE-D, 5 years ahead forecasts

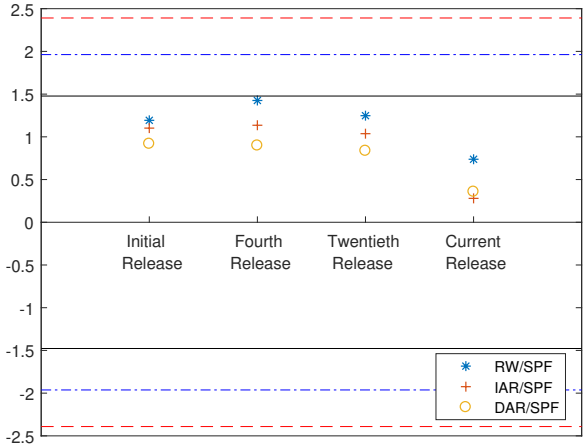
Figure 8: DM test statistic for unemployment and quadratic loss function. Sample 2002.Q1 - 2010.Q3. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.3911; blue dash-dotted: 10%, 1.9626; black solid: 20%, 1.4774) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.3986; blue dash-dotted: 10%, 1.9147; black solid: 20%, 1.4253).



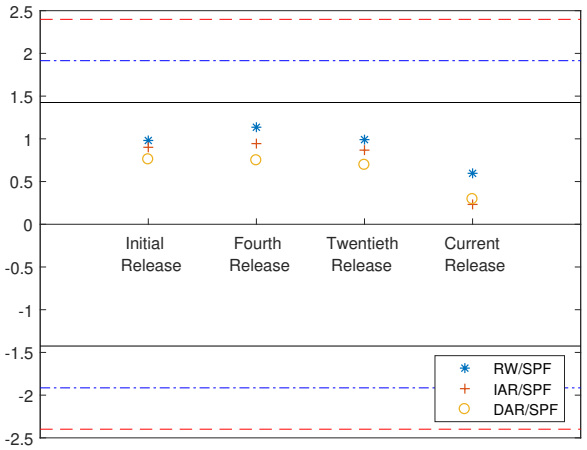
(a) WCE-B, 1 year ahead forecasts



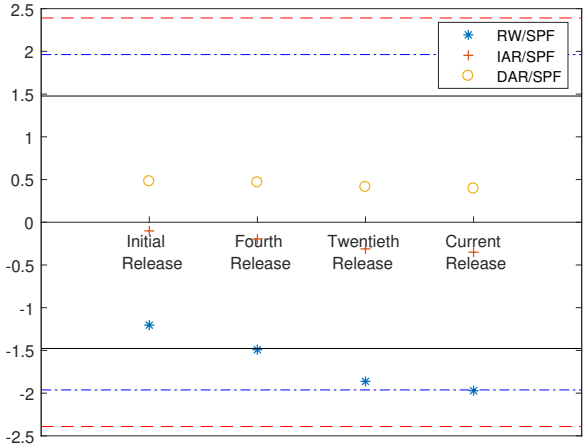
(b) WPE-D, 1 year ahead forecasts



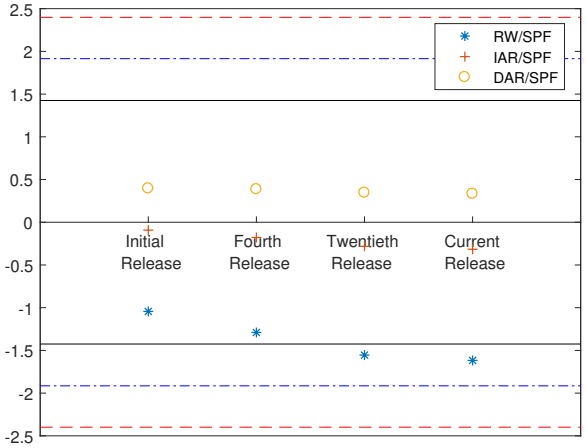
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

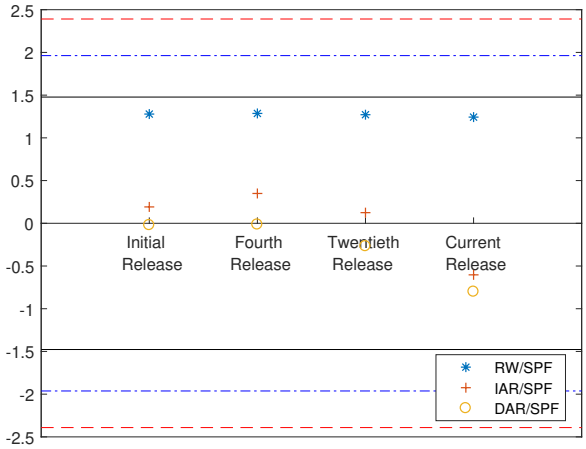


(e) WCE-B, 5 years ahead forecasts

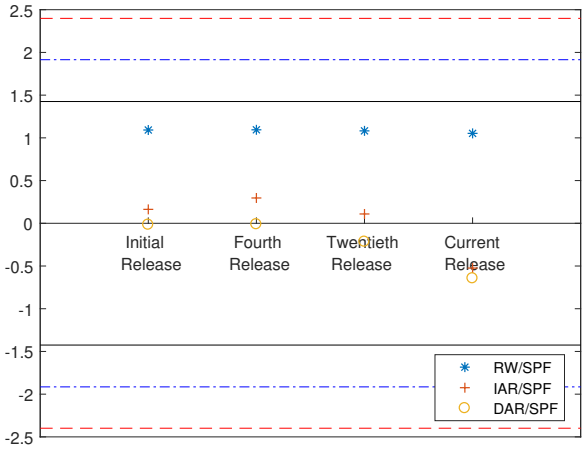


(f) WPE-D, 5 years ahead forecasts

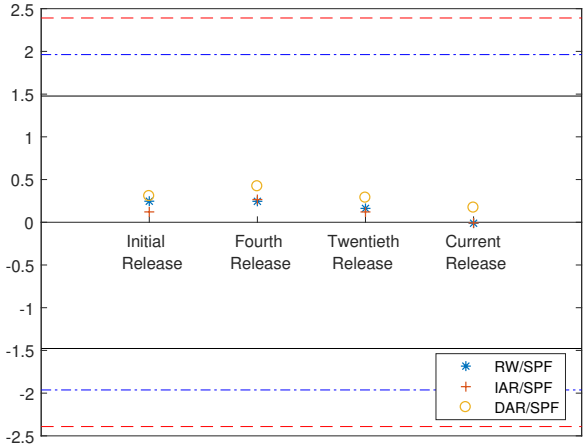
Figure 9: DM test statistic for unemployment and absolute loss function. Sample 2002.Q1 - 2010.Q3. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.3911; blue dash-dotted: 10%, 1.9626; black solid: 20%, 1.4774) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.3986; blue dash-dotted: 10%, 1.9147; black solid: 20%, 1.4253).



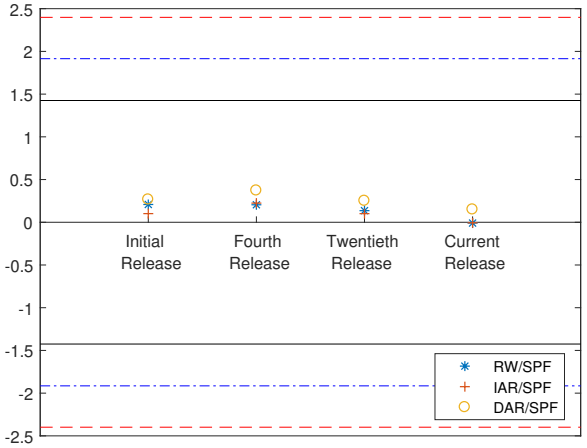
(a) WCE-B, 1 year ahead forecasts



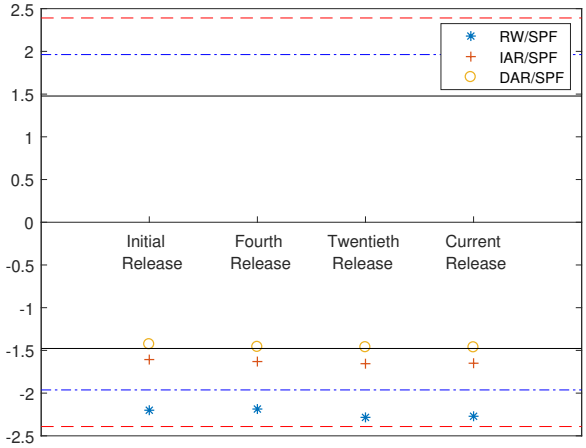
(b) WPE-D, 1 year ahead forecasts



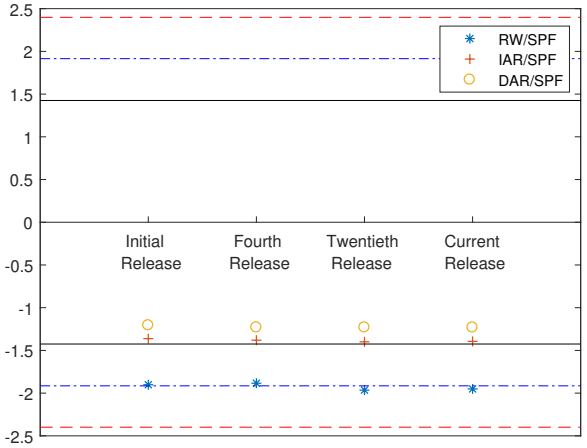
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

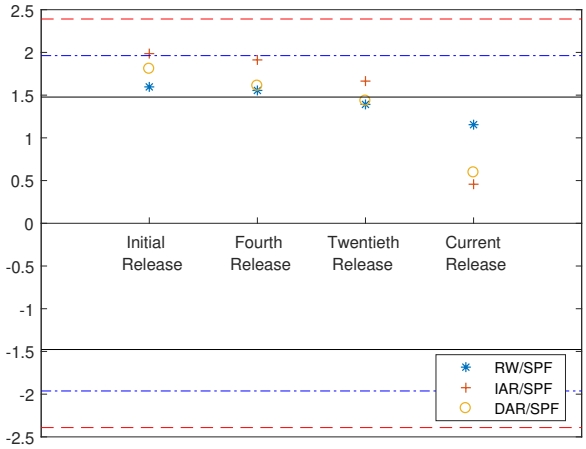


(e) WCE-B, 5 years ahead forecasts

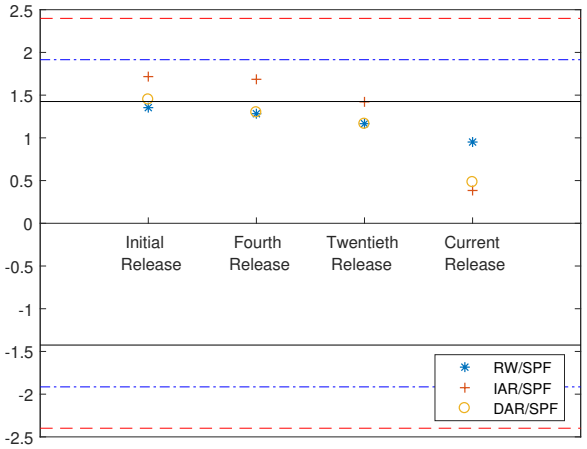


(f) WPE-D, 5 years ahead forecasts

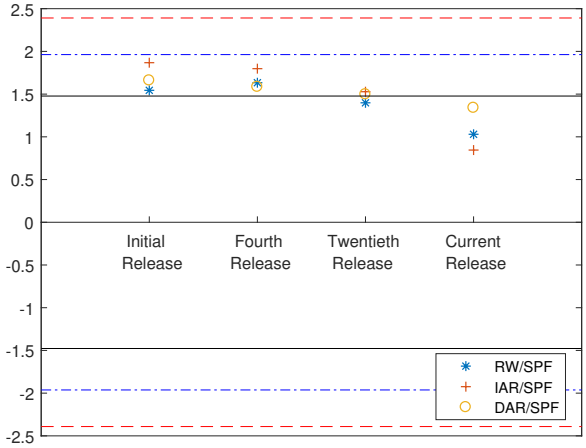
Figure 10: DM test statistic for unemployment and Linex loss function $\alpha = 1$. Sample 2002.Q1 - 2010.Q3. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.3911; blue dash-dotted: 10%, 1.9626; black solid: 20%, 1.4774) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.3986; blue dash-dotted: 10%, 1.9147; black solid: 20%, 1.4253).



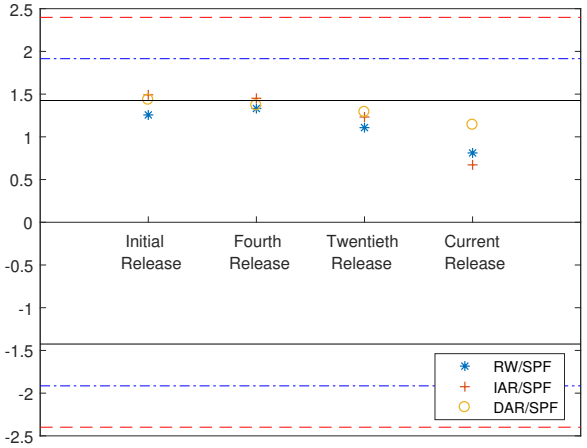
(a) WCE-B, 1 year ahead forecasts



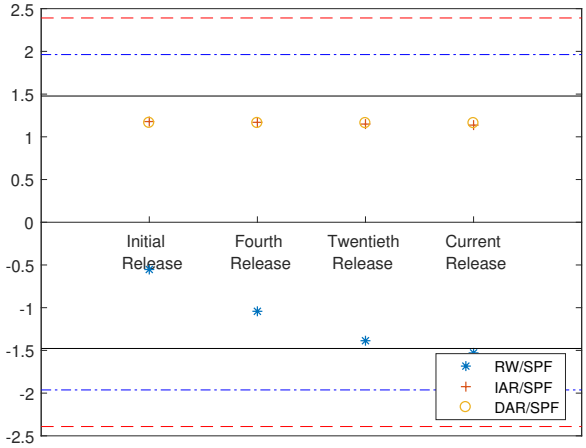
(b) WPE-D, 1 year ahead forecasts



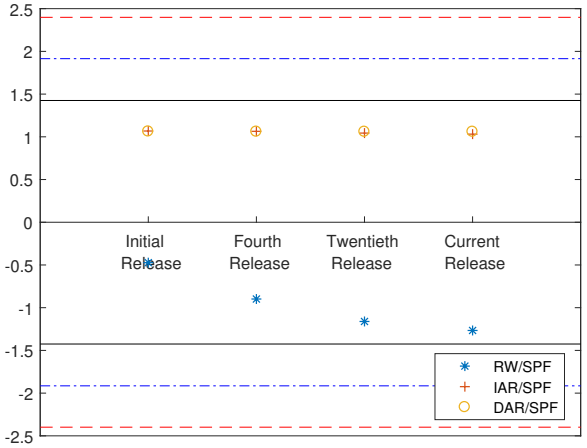
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

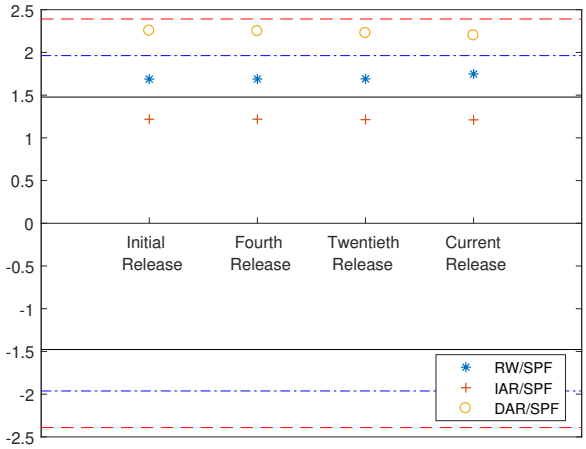


(e) WCE-B, 5 years ahead forecasts

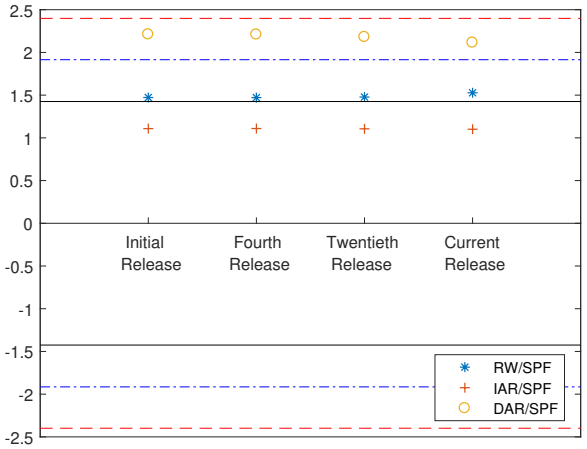


(f) WPE-D, 5 years ahead forecasts

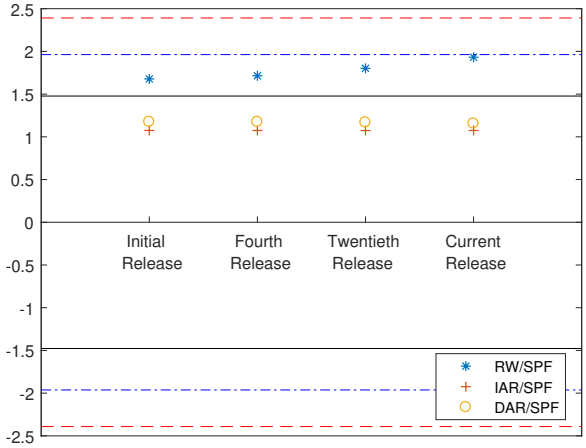
Figure 11: DM test statistic for unemployment and Linex loss function $\alpha = -1$. Sample 2002.Q1 - 2010.Q3. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.3911; blue dash-dotted: 10%, 1.9626; black solid: 20%, 1.4774) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.3986; blue dash-dotted: 10%, 1.9147; black solid: 20%, 1.4253).



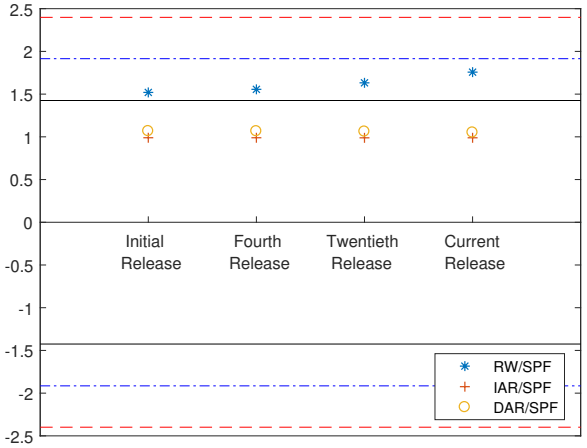
(a) WCE-B, 1 year ahead forecasts



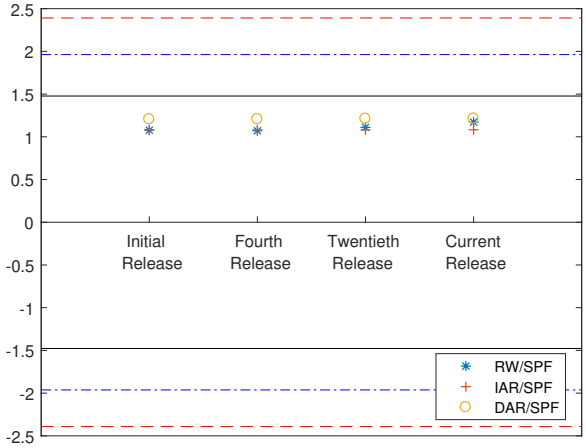
(b) WPE-D, 1 year ahead forecasts



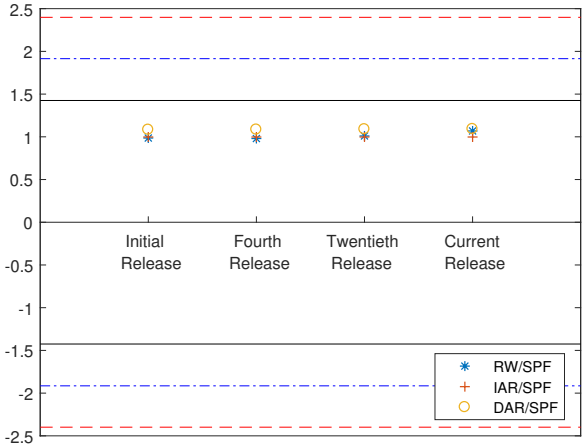
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

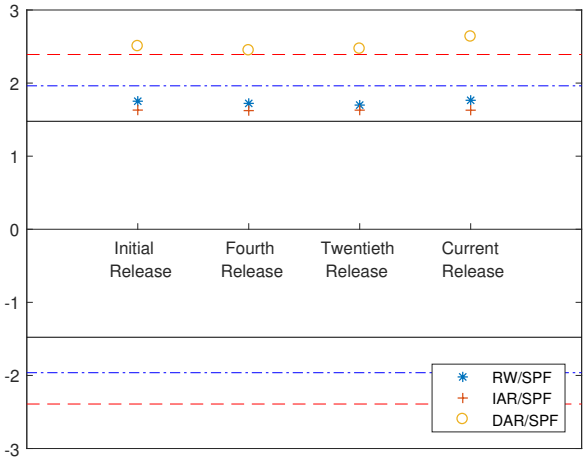


(e) WCE-B, 5 years ahead forecasts

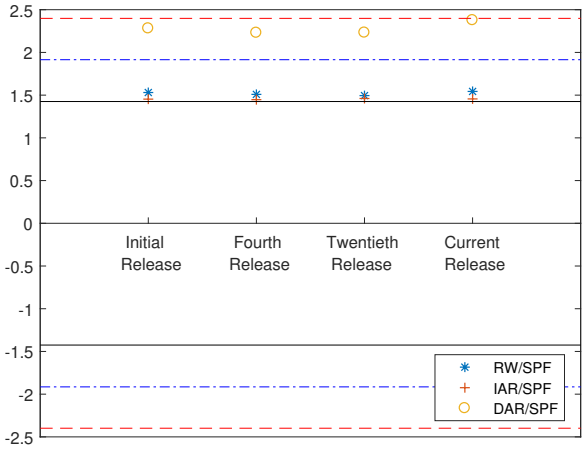


(f) WPE-D, 5 years ahead forecasts

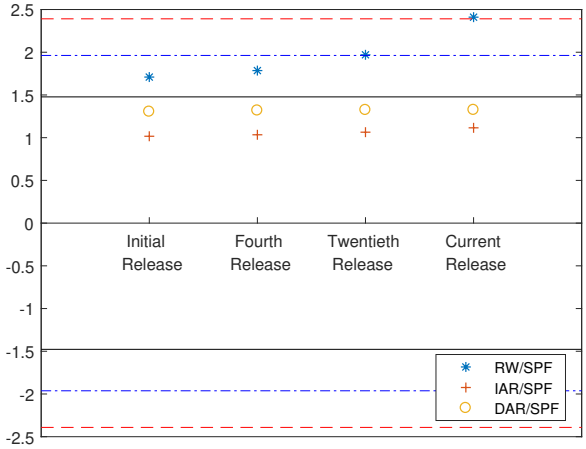
Figure 12: DM test statistic for real GDP growth and quadratic loss function. Sample 2002.Q1 - 2010.Q3. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.3911; blue dash-dotted: 10%, 1.9626; black solid: 20%, 1.4774) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.3986; blue dash-dotted: 10%, 1.9147; black solid: 20%, 1.4253).



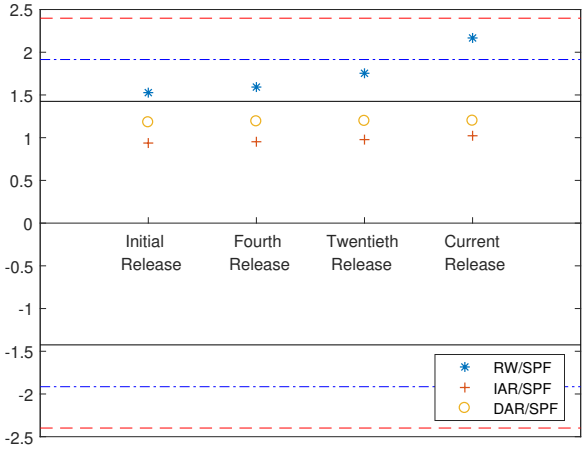
(a) WCE-B, 1 year ahead forecasts



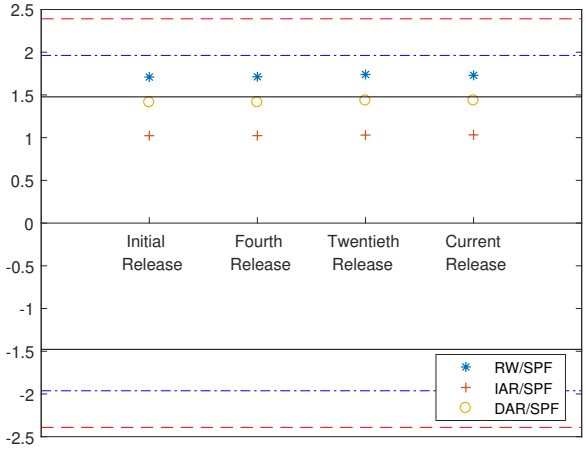
(b) WPE-D, 1 year ahead forecasts



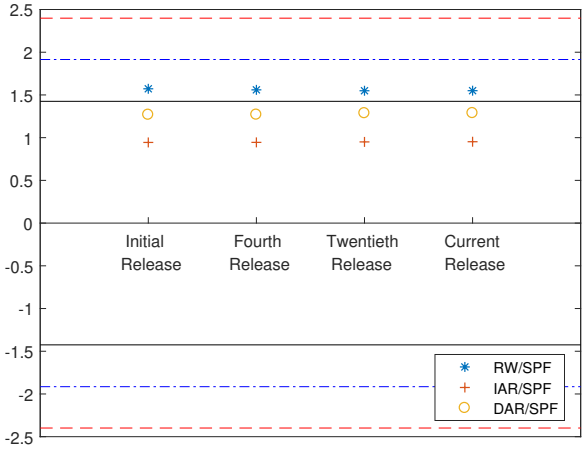
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

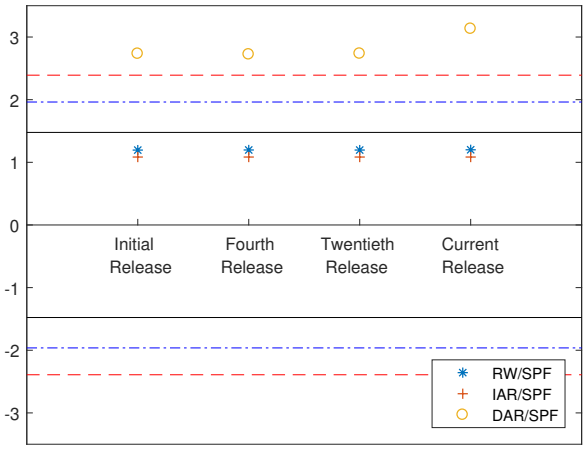


(e) WCE-B, 5 years ahead forecasts

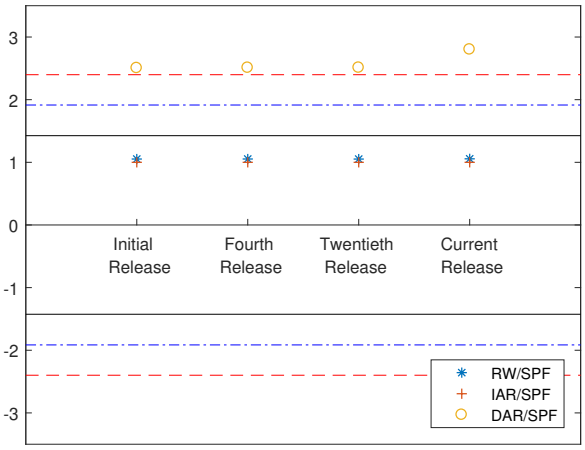


(f) WPE-D, 5 years ahead forecasts

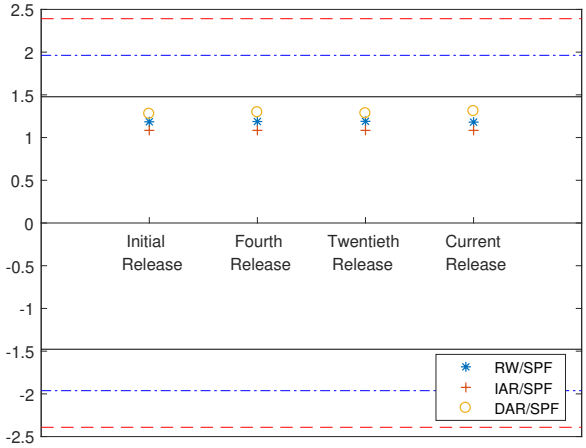
Figure 13: DM test statistic for real GDP growth and absolute loss function. Sample 2002.Q1 - 2010.Q3. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.3911; blue dash-dotted: 10%, 1.9626; black solid: 20%, 1.4774) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.3986; blue dash-dotted: 10%, 1.9147; black solid: 20%, 1.4253).



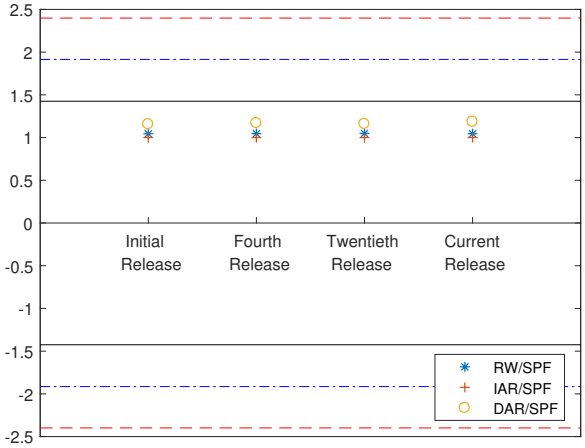
(a) WCE-B, 1 year ahead forecasts



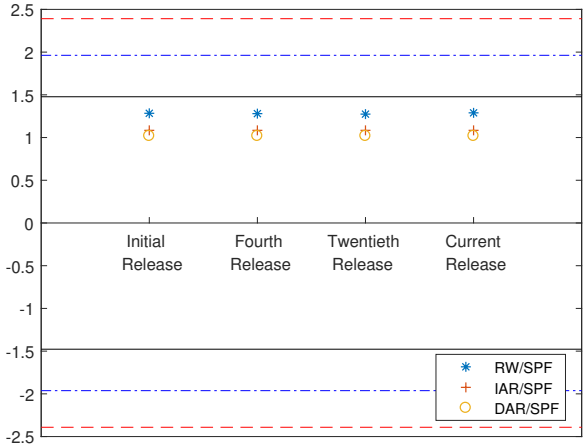
(b) WPE-D, 1 year ahead forecasts



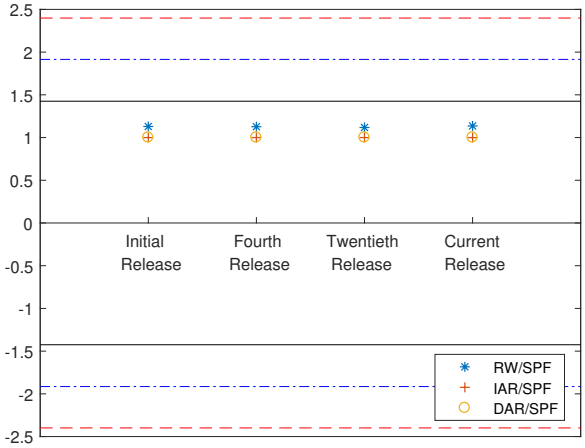
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

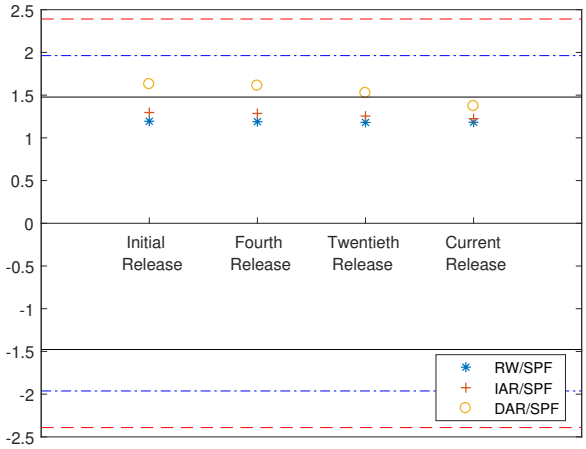


(e) WCE-B, 5 years ahead forecasts

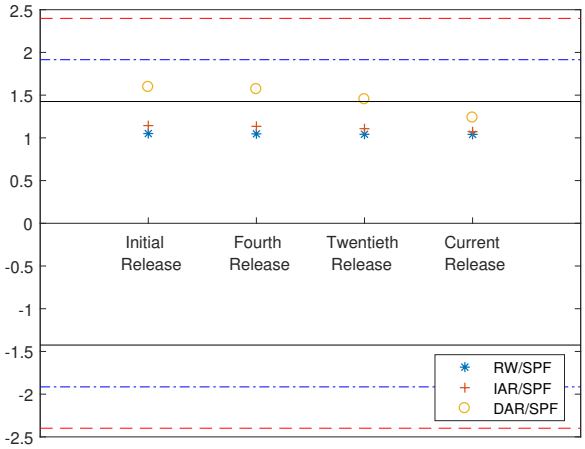


(f) WPE-D, 5 years ahead forecasts

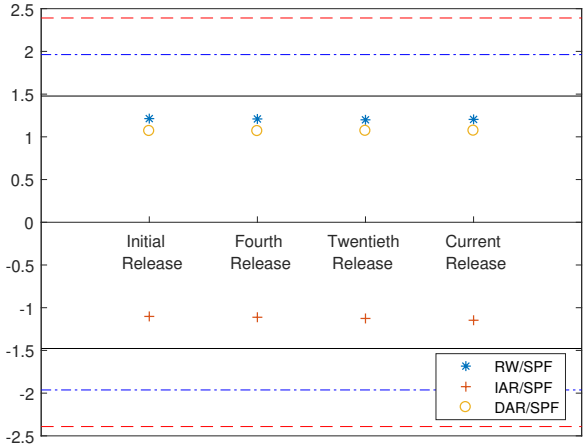
Figure 14: DM test statistic for real GDP growth and Linex loss function $\alpha = 1$. Sample 2002.Q1 - 2010.Q3. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.3911; blue dash-dotted: 10%, 1.9626; black solid: 20%, 1.4774) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.3986; blue dash-dotted: 10%, 1.9147; black solid: 20%, 1.4253).



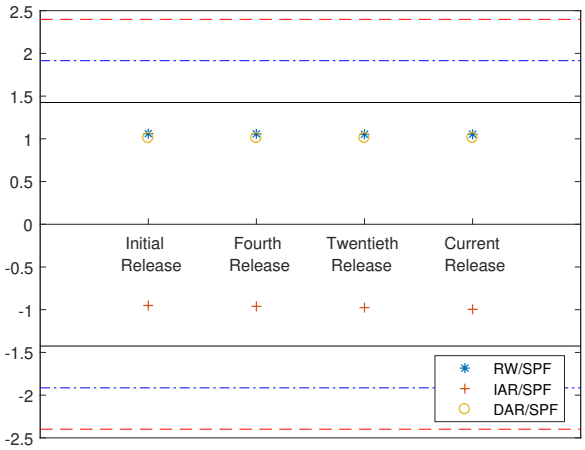
(a) WCE-B, 1 year ahead forecasts



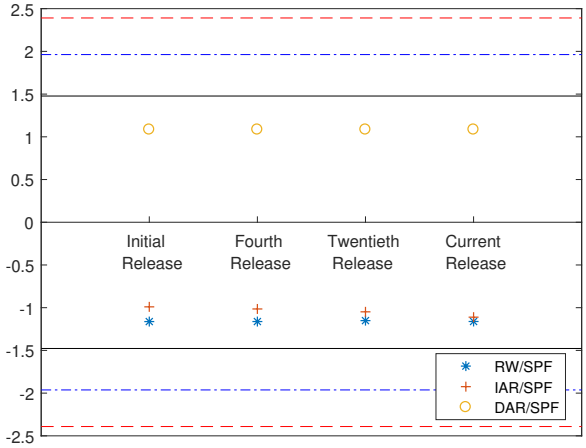
(b) WPE-D, 1 year ahead forecasts



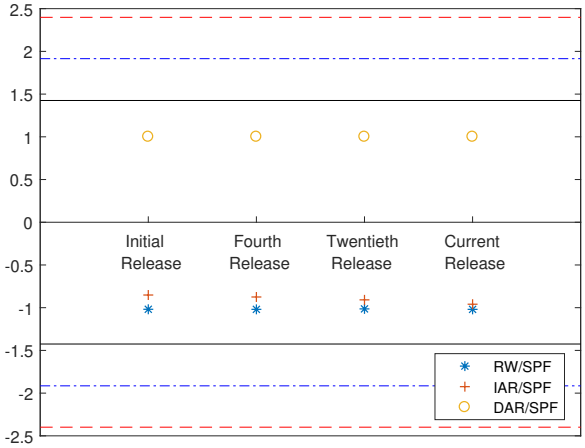
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

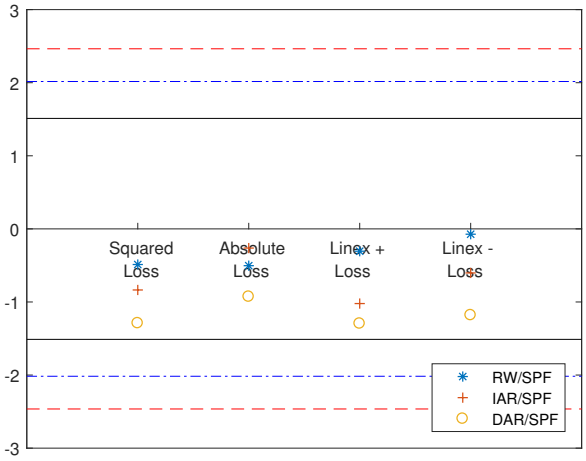


(e) WCE-B, 5 years ahead forecasts

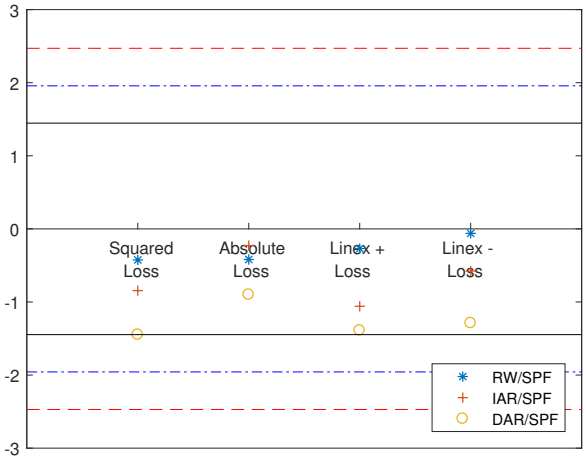


(f) WPE-D, 5 years ahead forecasts

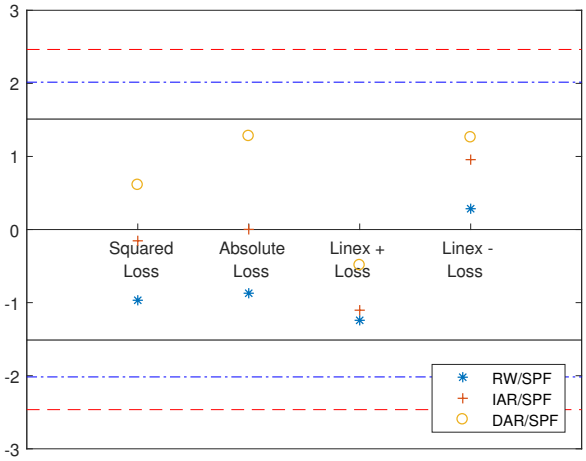
Figure 15: DM test statistic for real GDP growth and Linex loss function $\alpha = -1$. Sample 2002.Q1 - 2010.Q3. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.3911; blue dash-dotted: 10%, 1.9626; black solid: 20%, 1.4774) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.3986; blue dash-dotted: 10%, 1.9147; black solid: 20%, 1.4253).



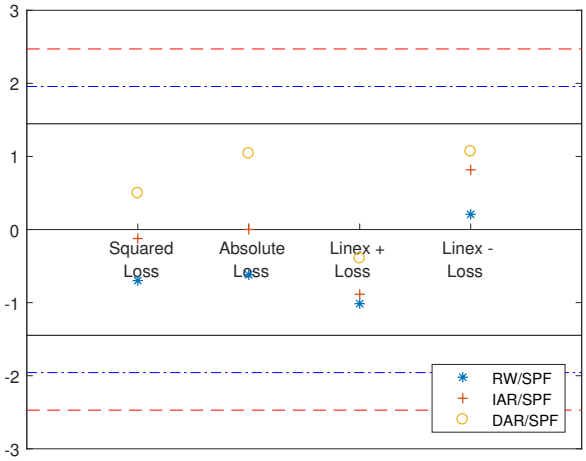
(a) WCE-B, 1 year ahead forecasts



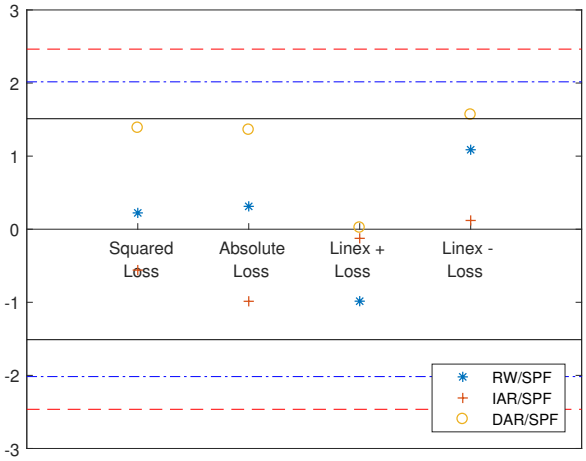
(b) WPE-D, 1 year ahead forecasts



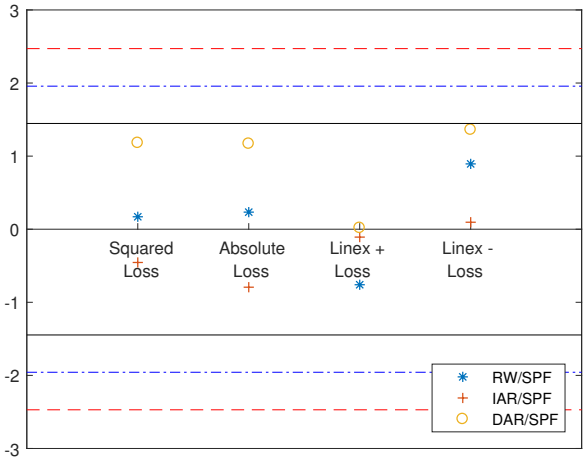
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

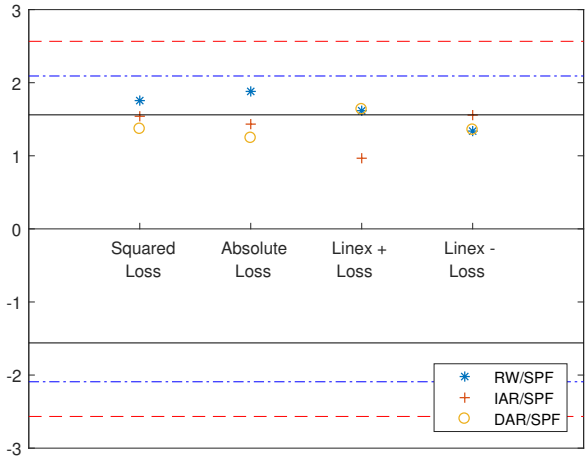


(e) WCE-B, 5 years ahead forecasts

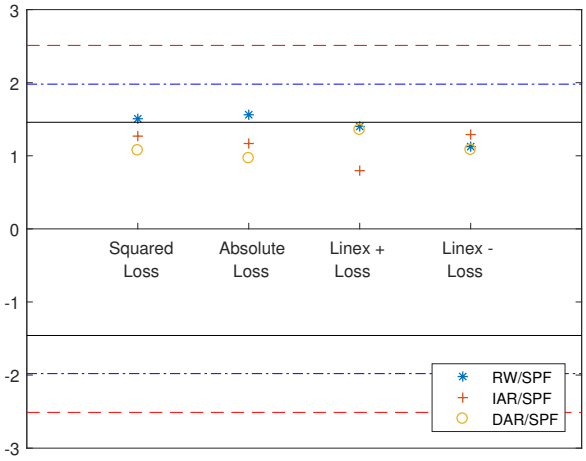


(f) WPE-D, 5 years ahead forecasts

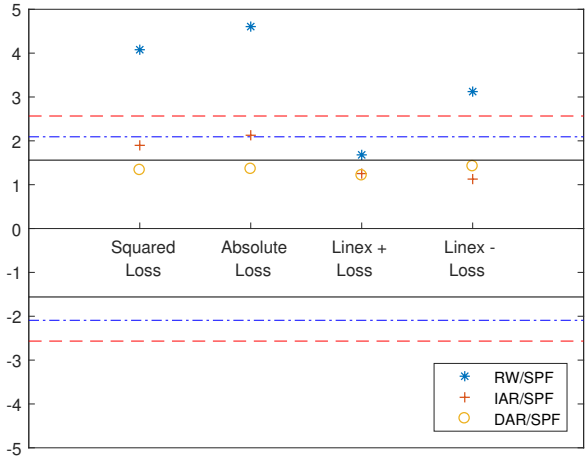
Figure 16: DM test statistic for inflation. Sub sample 2002.Q1 - 2007.Q4. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.4640; blue dash-dotted: 10%, 2.0164; black solid: 20%, 1.5116) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.4709; blue dash-dotted: 10%, 1.9572; black solid: 20%, 1.4469).



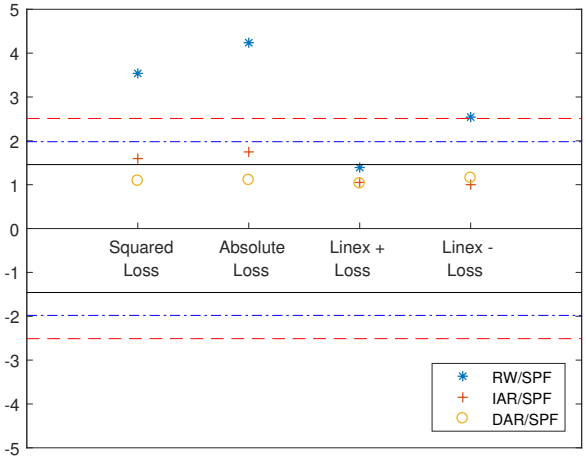
(a) WCE-B, 1 year ahead forecasts



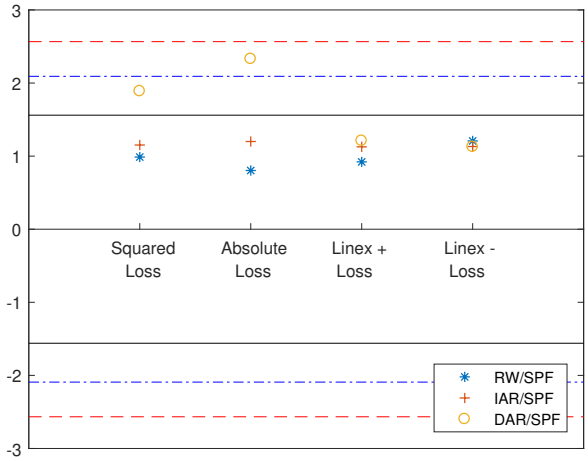
(b) WPE-D, 1 year ahead forecasts



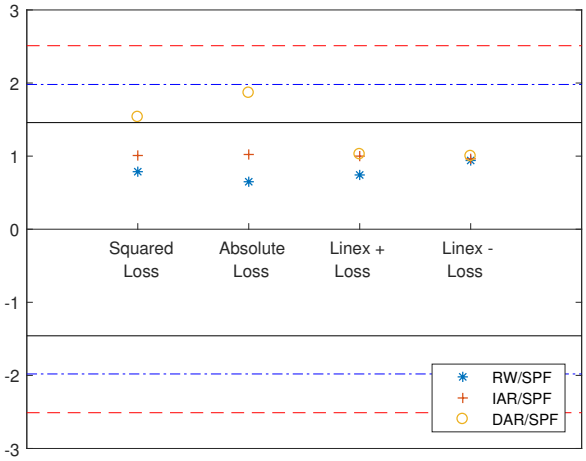
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

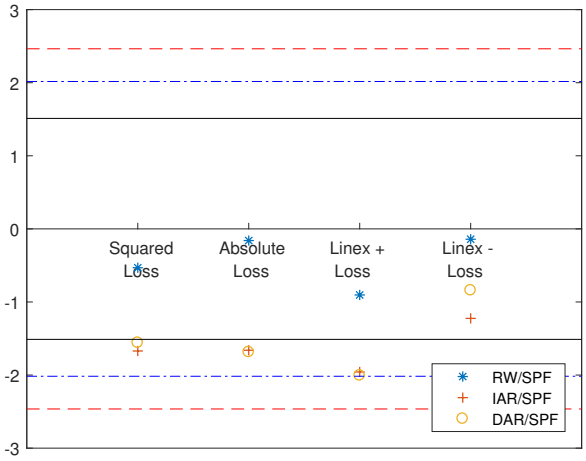


(e) WCE-B, 5 years ahead forecasts

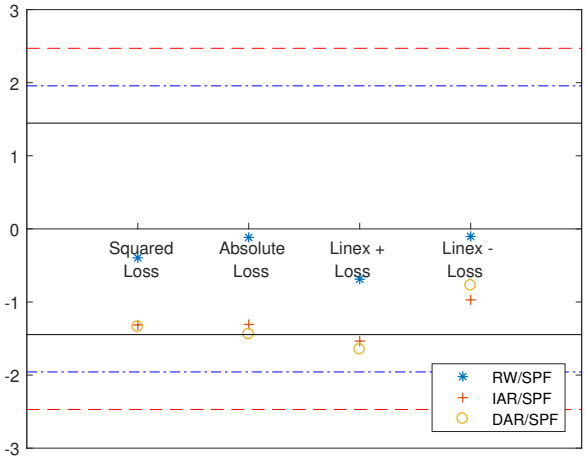


(f) WPE-D, 5 years ahead forecasts

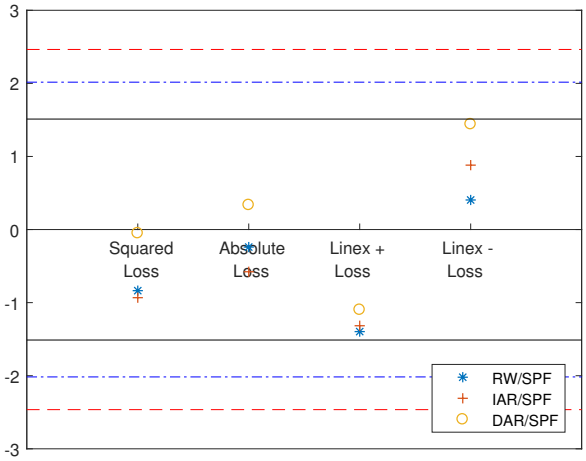
Figure 17: DM test statistic for inflation. Sub sample 2008.Q1 - 2012.Q4. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.5663; blue dash-dotted: 10%, 2.0919; black solid: 20%, 1.5602) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.5107; blue dash-dotted: 10%, 1.9804; black solid: 20%, 1.4586).



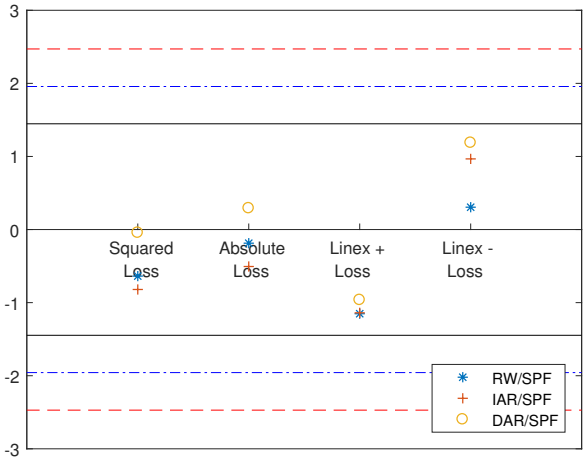
(a) WCE-B, 1 year ahead forecasts



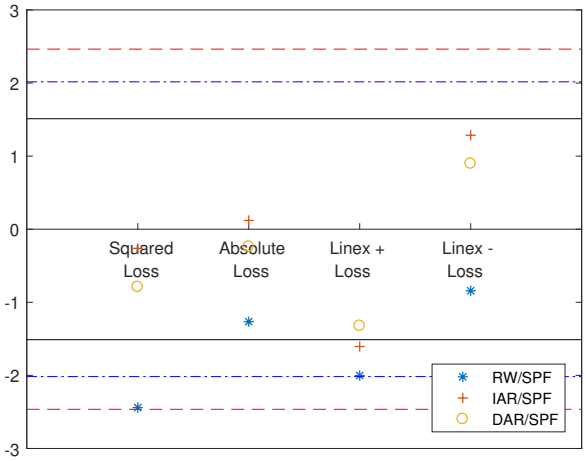
(b) WPE-D, 1 year ahead forecasts



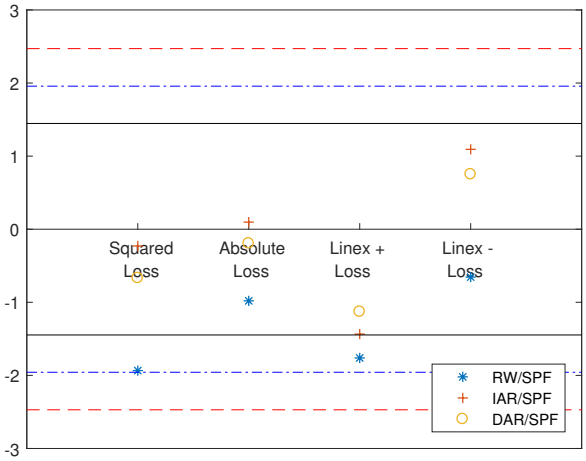
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

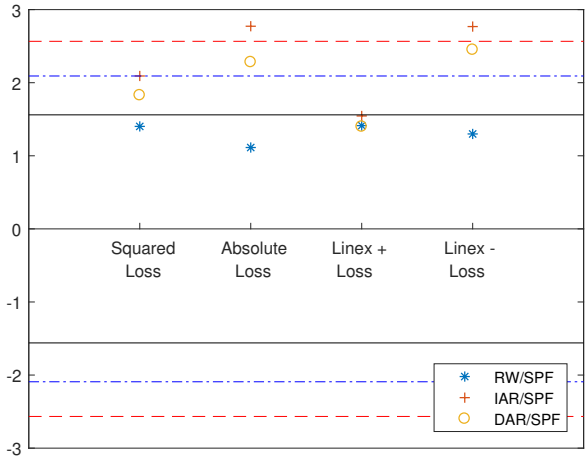


(e) WCE-B, 5 years ahead forecasts

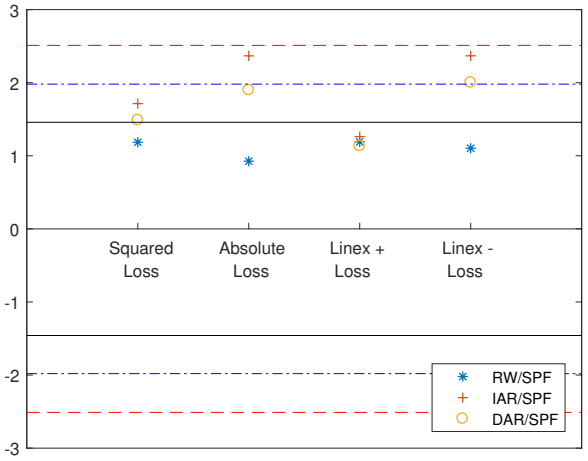


(f) WPE-D, 5 years ahead forecasts

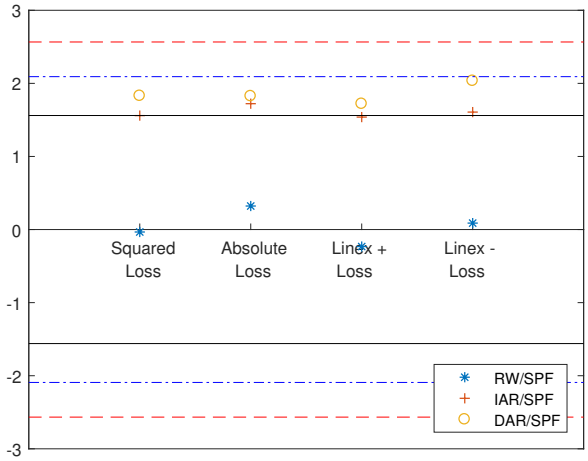
Figure 18: DM test statistic for unemployment. Sub sample 2002.Q1 - 2007.Q4. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.4640; blue dash-dotted: 10%, 2.0164; black solid: 20%, 1.5116) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.4709; blue dash-dotted: 10%, 1.9572; black solid: 20%, 1.4469).



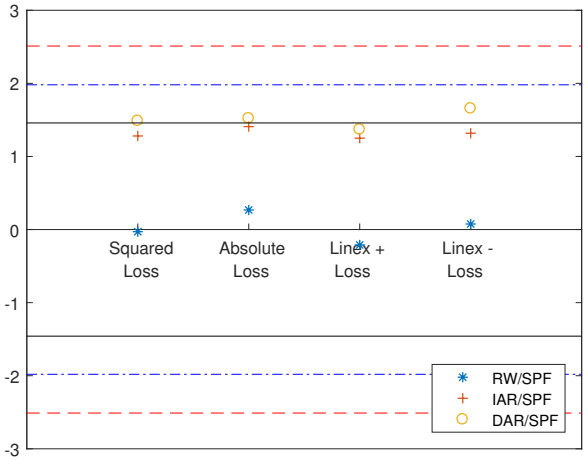
(a) WCE-B, 1 year ahead forecasts



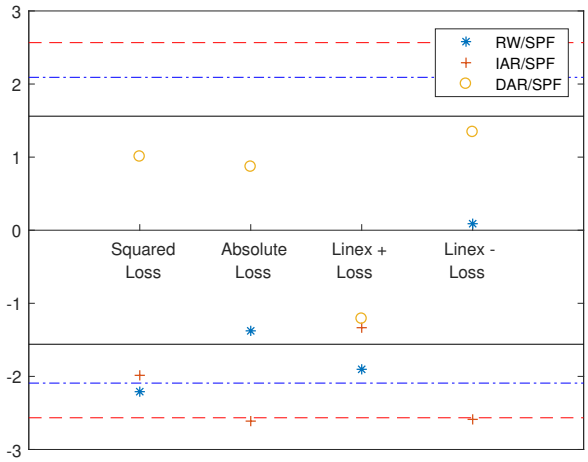
(b) WPE-D, 1 year ahead forecasts



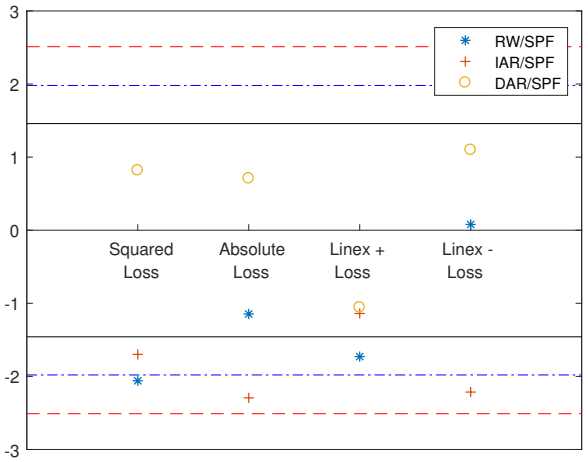
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

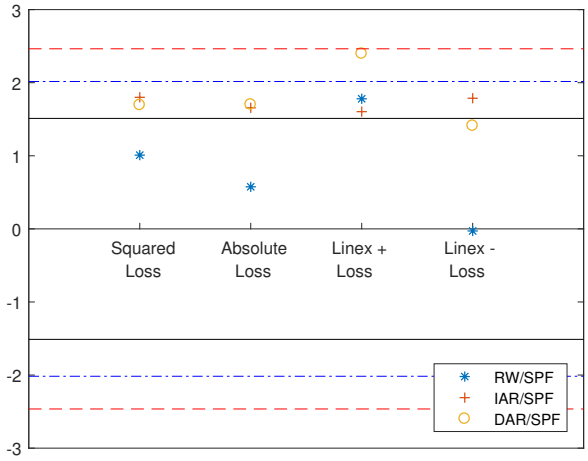


(e) WCE-B, 5 years ahead forecasts

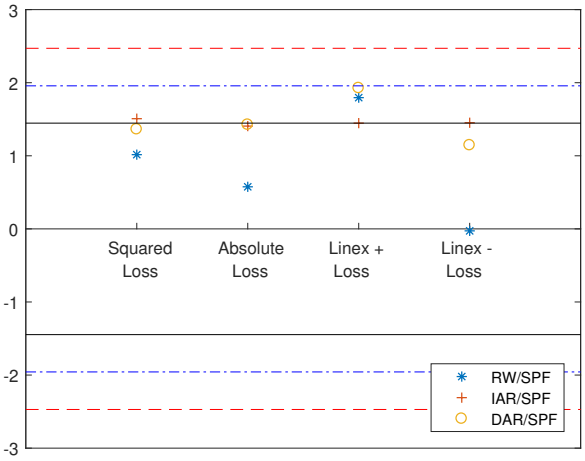


(f) WPE-D, 5 years ahead forecasts

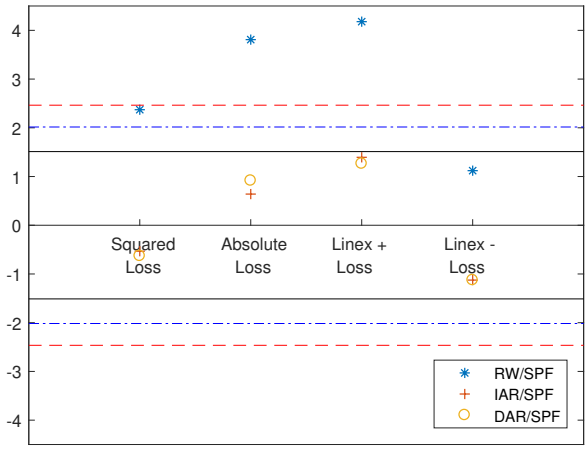
Figure 19: DM test statistic for unemployment. Sub sample 2008.Q1 - 2012.Q4. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.5663; blue dash-dotted: 10%, 2.0919; black solid: 20%, 1.5602) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.5107; blue dash-dotted: 10%, 1.9804; black solid: 20%, 1.4586).



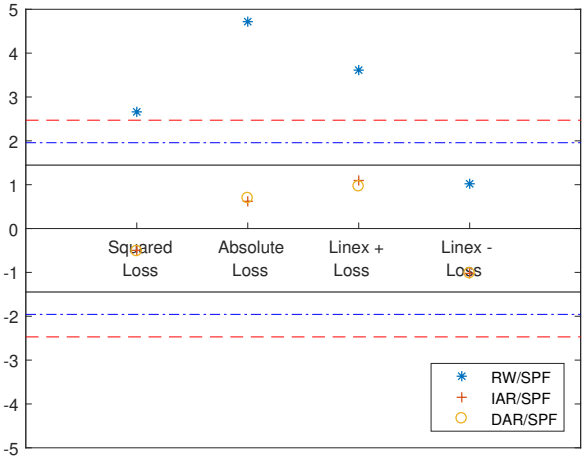
(a) WCE-B, 1 year ahead forecasts



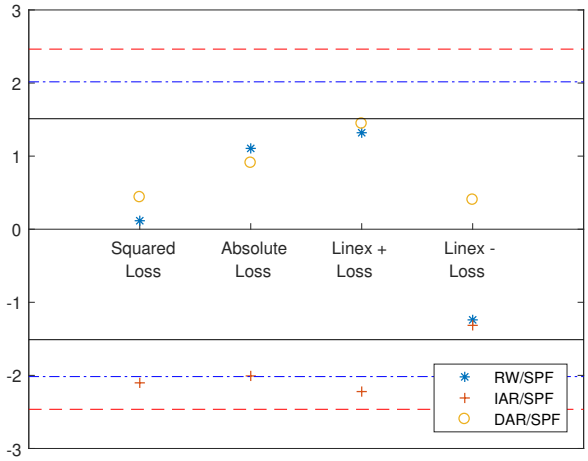
(b) WPE-D, 1 year ahead forecasts



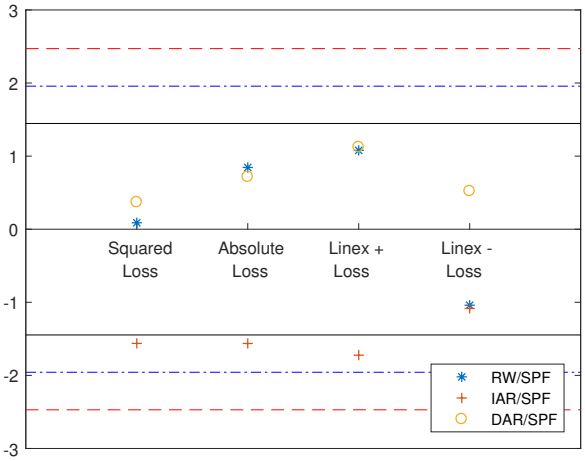
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts

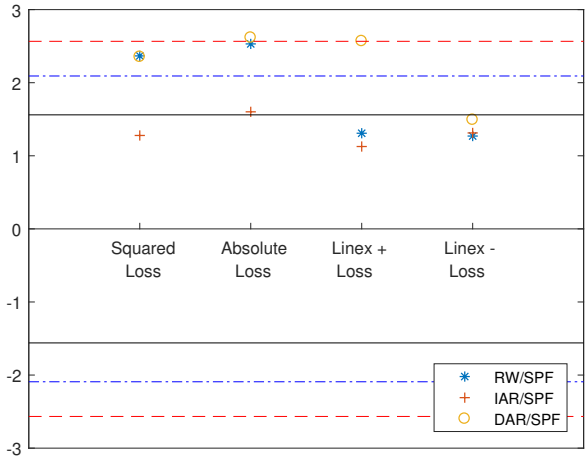


(e) WCE-B, 5 years ahead forecasts

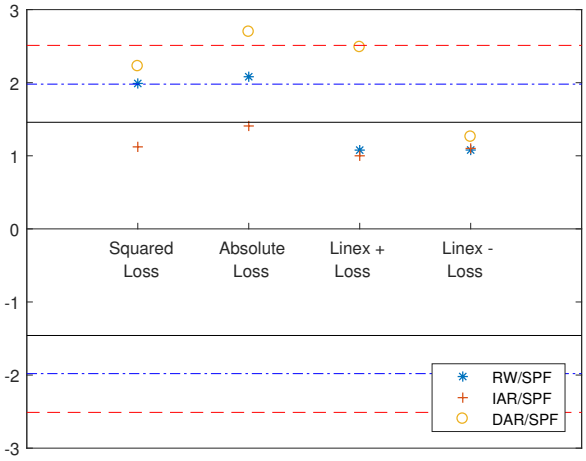


(f) WPE-D, 5 years ahead forecasts

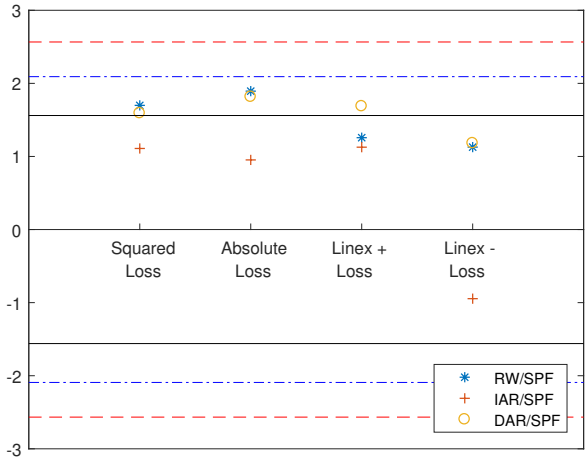
Figure 20: DM test statistic for real GDP growth. Sub sample 2002.Q1 - 2007.Q4. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.4640; blue dash-dotted: 10%, 2.0164; black solid: 20%, 1.5116) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.4709; blue dash-dotted: 10%, 1.9572; black solid: 20%, 1.4469).



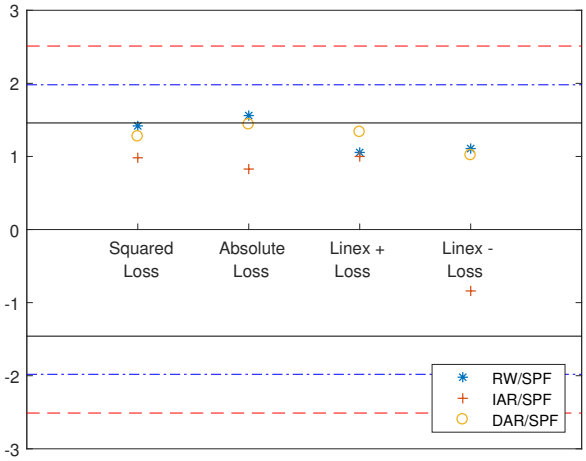
(a) WCE-B, 1 year ahead forecasts



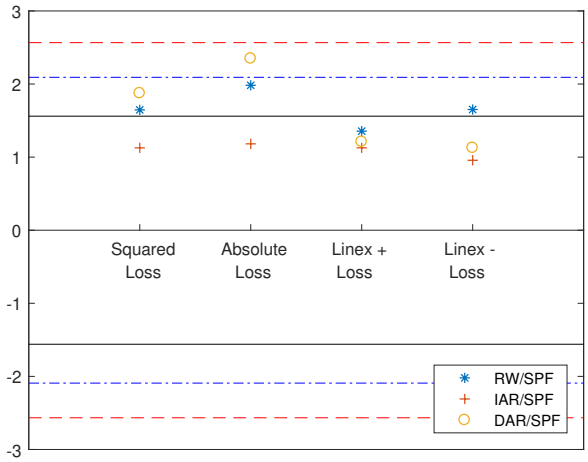
(b) WPE-D, 1 year ahead forecasts



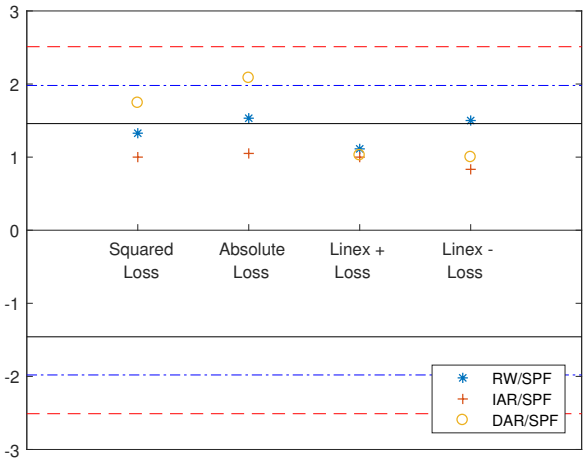
(c) WCE-B, 2 years ahead forecasts



(d) WPE-D, 2 years ahead forecasts



(e) WCE-B, 5 years ahead forecasts



(f) WPE-D, 5 years ahead forecasts

Figure 21: DM test statistic for real GDP growth. Sub sample 2008.Q1 - 2012.Q4. Lines are two side critical values taken from a non standard distribution in the case of WCE with fixed b asymptotics (red dashed: 5%, 2.5663; blue dash-dotted: 10%, 2.0919; black solid: 20%, 1.5602) and from a Student-t distribution with $2m$ degrees of freedom in the case of WPE with fixed m asymptotics (red dashed: 5%, 2.5107; blue dash-dotted: 10%, 1.9804; black solid: 20%, 1.4586).

References

Atkeson, A. and Ohanian, L. E. (2001). Are phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis. Quarterly Review-Federal Reserve Bank of Minneapolis*, *25*(1), 2.

Balcilar, M., Gupta, R., Majumdar, A., and Miller, S. M. (2015). Was the recent downturn in us real gdp predictable? *Applied Economics*, *47*(28), 2985–3007.

Bhansali, R. J. (2002). Multi-step forecasting. In *A Companion to Economic Forecasting* chapter 9, (pp. 206–221). Wiley.

Boero, G., Smith, J., and Wallis, K. F. (2008). Evaluating a three-dimensional panel of point forecasts: the bank of england survey of external forecasters. *International Journal of Forecasting*, *24*(3), 354–367.

Bowles, C., Friz, R., Genre, V., Kenny, G., Meyler, A., and Rautanen, T. (2007). The ecb survey of professional forecasters (spf)-a review after eight years' experience. *ECB Occasional Paper*, (50).

Bowles, C., Friz, R., Genre, V., Kenny, G., Meyler, A., and Rautanen, T. (2011). An evaluation of the growth and unemployment forecasts in the ecb survey of professional forecasters. *OECD Journal: Journal of Business Cycle Measurement and Analysis*, *2010*(2), 1–28.

Capistrán, C. and Timmermann, A. (2009). Disagreement and biases in inflation expectations. *Journal of Money, Credit and Banking*, *41*(2-3), 365–396.

Chong, Y. Y. and Hendry, D. F. (1986). Econometric evaluation of linear macro-economic models. *The Review of Economic Studies*, *53*(4), 671–690.

Clark, T. E. (1999). Finite-sample properties of tests for equal forecast accuracy. *Journal of Forecasting*, *18*(7), 489–504.

Clark, T. E. and McCracken, M. W. (2009). Tests of equal predictive ability with real-time data. *Journal of Business and Economic Statistics*, *27*(4), 441–454.

Clements, M. P. (2009). Internal consistency of survey respondents' forecasts: Evidence based on the survey of professional forecasters. In *The methodology and practice of econometrics. A festschrift in honour of David F. Hendry* (pp. 206–226). Oxford University Press Oxford.

Clements, M. P. (2010). Explanations of the inconsistencies in survey respondents' forecasts. *European Economic Review*, *54*(4), 536–549.

Clements, M. P. and Galvão, A. B. (2012). Improving real-time estimates of output and inflation gaps with multiple-vintage models. *Journal of Business and Economic Statistics*, *30*(4), 554–562.

Coroneo, L. and Iacone, F. (2015). Comparing predictive accuracy in small samples. Technical Report 15/15, The University of York.

Croushore, D. (2006). Forecasting with real-time macroeconomic data. *Handbook of economic forecasting*, *1*, 961–982.

Croushore, D. and Stark, T. (2001). A real-time data set for macroeconomists. *Journal of econometrics*, *105*(1), 111–130.

D'Agostino, A., Giannone, D., and Surico, P. (2006). (un) predictability and macroeconomic stability. *ECB Working Paper Series*, (605).

Demetrescu, M., Hanck, C., and Kruse, R. (2018). Robust inference under time-varying volatility: A real-time evaluation of professional forecasters. Technical report.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and economic statistics*, *13*(3), 253–262.

Elliott, G., Komunjer, I., and Timmermann, A. (2008). Biases in macroeconomic forecasts: irrationality or asymmetric loss? *Journal of the European Economic Association*, *6*(1), 122–157.

Engelberg, J., Manski, C. F., and Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business and Economic Statistics*, *27*(1), 30–41.

Fair, R. C. and Shiller, R. J. (1989). The informational content of ex ante forecasts. *The Review of Economics and Statistics*, *71*(2), 325–331.

Fair, R. C. and Shiller, R. J. (1990). Comparing information in forecasts from econometric models. *The American Economic Review*, *80*(3), 375–389.

Garcia, J. A. (2003). An introduction to the ecb's survey of professional forecasters. *ECB Occasional Paper*, (8).

Garcia, J. A. and Manzanares, A. (2007). Reporting biases and survey results: evidence from european professional forecasters. *ECB Working Paper Series*, (836).

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*(6), 1545–1578.

Giannone, D., Henry, J., Lalik, M., and Modugno, M. (2012). An area-wide real-time database for the euro area. *Review of Economics and Statistics*, *94*(4), 1000–1013.

Granger, C. W. (1999). Outline of forecast theory using generalized cost functions. *Spanish Economic Review*, *1*(2), 161–173.

Harvey, D. I., Leybourne, S. J., and Whitehouse, E. J. (2016). Testing forecast accuracy in small samples. Technical report, School of Economic University of Nottingham.

Hualde, J. and Iacone, F. (2015). Autocorrelation robust inference using the daniell kernel with fixed bandwidth. Technical Report 15/14, The University of York.

Kiefer, N. M. and Vogelsang, T. J. (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, *21*(6), 1130–1164.

Mankiw, N. G. and Shapiro, M. D. (1986). News or noise? an analysis of gnp revisions.

Marcellino, M., Stock, J. H., and Watson, M. W. (2006). A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of econometrics*, *135*(1-2), 499–526.

Mincer, J. A. and Zarnowitz, V. (1969). The evaluation of economic forecasts. In *Economic forecasts and expectations: Analysis of forecasting behavior and performance* (pp. 3–46). NBER.

Schorfheide, F. (2005). Var forecasting under misspecification. *Journal of Econometrics*, *128*(1), 99–136.

Stark, T. (2010). Realistic evaluation of real-time forecasts in the survey of professional forecasters. *Federal Reserve Bank of Philadelphia Research Rap, Special Report*, *1*.

Stark, T. and Croushore, D. (2002). Forecasting with a real-time data set for macroeconomists. *Journal of Macroeconomics*, *24*(4), 507–531.

Theil, H. (1958). *Economic forecasts and policy*. North-Holland.

Varian, H. R. (1975). A bayesian approach to real estate assessment. *Studies in Bayesian Econometric and Statistics in honor of Leonard J. Savage*, *81*(394), 195–208.

Wilson, E. B. (1934). The periodogram of american business activity. *The Quarterly Journal of Economics*, *48*(3), 375–417.