

New Methods for Comparing the Forecasts Accuracy

Bratu (Simionescu) Mihaela *

Department of Statistics and Econometrics, Faculty of Cybernetics, Statistics and Economic Informatics, Bucharest, Romania

*Corresponding author: mihaela_mb1@yahoo.com

Received January 08, 2013; Revised January 28, 2013; Accepted February 28, 2013

Abstract The main purpose of this research is to show the diversity of statistical methods that could be used to assess and compare forecasts accuracy. Some of the statistical approaches were not used before in literature to evaluate the forecasts accuracy. The different methods applied to compare the accuracy of the USA inflation forecasts on the horizon 1976-2012 started from the predictions provided by Survey of Professional Forecasters (SPF), Congressional Budget Office (CBO), Blue Chips (BC), and Administration, determining different results. According to U1 Theil's statistic, non-parametric tests and a new indicator proposed by us (RRSSE- ratio of radicals of sum of squared errors), the best forecasts were provided by Administration and the less accurate by SPF. The Spearman's and Kendall's coefficients of correlation and the ranks method gave a hierarchy of institutions performance regarding the accuracy that starts with BC and finished with SPF. The logistic regression computed by the author and the relative distance to the maximal performance method considered CBO as the best institution. Some methods of improving the forecasts accuracy were applied, getting more accurate predictions for the combined forecasts of BC and CBO using optimal scheme of combination. The smoothed predicted values based on Hodrick-Prescott filter outperformed all the initial predictions and the combined ones.

Keywords: forecasts, accuracy, logistic regression, combined forecasts, non-parametric tests, filters, multi-criteria ranking

1. Introduction

An auxiliary, but essential component of the forecasting process is the assessment of the accuracy, which reflects how closer the forecasted values of a variable are to its registered values. In USA there are more institutions that provide predictions for macroeconomic indicators. The main question is which of these institutions predicted the best an economic phenomenon. To answer this question we can use many methods. Some of the usual statistical methods were not applied in the forecasting context. It is important to analyze if more methods gave the same results. On the other hand, it is also essential to find out some empirical strategies to improve the forecasts accuracy.

In this study the accuracy is assessed in ex-post variant, resulting a mirror of the historical accuracy of institutions forecasts. Consequently, this analyze will be the best guide to choose the forecasts of a certain institution in the close future.

2. New Statistical Methods Used in Making Comparisons between Predictions

Some researchers were interested in evaluating the accuracy of macroeconomic forecasts made by some international institutions. However, they omitted to take into account the Administration anticipations.

Edge, Kiley and Laforte assessed the accuracy of predictions made by Federal Reserve staff and for those made starting from a DSGE model and a time-series model [7].

Abreu were interested in assessing the accuracy of predictions made by the Organization for Economic Cooperation and Development (OECD), International Monetary Fund (IMF), European Commission (EC) and two private institutions (Consensus Economics and The Economist) [1]. The directional accuracy and the ability of anticipating an economic crisis were deeply studied. The probability of USA recession was computed by Österholm (2012) starting from a BVAR model [9].

In general, the researchers used the classical measures of accuracy, like mean error, mean absolute errors, root mean squared error, U Theil's coefficient. Percentages errors, U statistics or mean absolute scaled errors are used to make comparisons in terms of accuracy.

One problem that we try to solve in this study is to bring an objective classification of forecasts. In practice, some of the accuracy measures recommend as better a certain forecast, while others show that other forecast is more accurate. In order to solve this uncertain situation we applied the multi-criteria methods that take into account at the same time the values of all accuracy indicators.

The ranks method and the method of relative distance with respect to the maximal performance are applied to order the forecasts according to accuracy criterion.

For the forecasted variable X , the error is calculated as the difference between the real value and the predicted one. It is denoted by "e". Some of the measures of predictions accuracy are presented below. We selected 5

accuracy measures whose influence is taken into account at the same time.

If n is the length of the forecast horizon, then we computed:

1. The Mean error (ME)

$$ME = \frac{1}{n} \sum_{i=1}^n e_x \tag{1}$$

2. The Mean absolute error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |e_x| \tag{2}$$

3. The Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n e_x^2} \tag{3}$$

4. The U1 Theil's statistic

$$U_1 = \frac{\sqrt{\sum_{t=1}^n (r_t - f_t)^2}}{\sqrt{\sum_{t=1}^n r_t^2} + \sqrt{\sum_{t=1}^n f_t^2}} \tag{4}$$

r - the real values; f - the forecasted values

A higher accuracy is equivalent with a value closer to zero for U1 statistic.

5. U2 Theil's statistic

$$U_2 = \sqrt{\frac{\sum_{t=1}^{n-1} \left(\frac{r_{t+1} - r_t}{r_t}\right)^2}{\sum_{t=1}^{n-1} \left(\frac{r_{t+1} - r_t}{r_t}\right)^2}} \tag{5}$$

A value less than 1 for U2 confirms the superiority of the compared forecast, while a value greater than 1 shows a higher accuracy for the benchmark forecast.

For comparisons with the naive forecasts a new indicator is introduced by us: ratio of radicals of sum of squared errors (RRSSE).

$$RRSSE = \frac{\sqrt{\sum_{t=1}^n e_t^2}}{\sqrt{\sum_{i=2}^{n-1} (x_i - x_{i-1})^2}}$$

In order to compare two predictions even for different variables, the values of this indicator are compared, a value closer to zero showing a better accuracy.

Ranks method has several steps:

1. Each accuracy measure receives a rank according to its value. (the value that indicates the a better degree accuracy has the rank 1);

The statistical units are represented by the number of institutions that provided forecasts. In our case study this number is 3. The rank corresponding to each institution is:

$$(r_{i_{ind_j}})_{i=1,2,3,4}$$

and ind_j -accuracy indicator j .

The ranks are sum up and institution with the lowest score receives the rank one:

$$S_i = \sum_{j=1}^5 (r_{i_{ind_j}}), i = 1, 2, 3, 4 \tag{6}$$

2. The institution with the lowest score is the best one and it gets the final rank 1.

The method of relative distance related to the maximal performance supposes that for each accuracy measure the distance of each institution compared to the one with the best performance is calculated as:

$$d_{i_{ind_j}} = \frac{ind_i^j}{\{\min abs(ind_i^j)\}_{i=1,\dots,4}}, \tag{7}$$

$i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4, 5$.

The relative distance is calculated as a ratio, where the denominator is the lowest value of the accuracy indicator for all institutions.

The geometric mean for the relative distances is calculated:

$$\bar{d}_i = \sqrt[5]{\prod_{j=1}^5 d_{i_{ind_j}}}, \tag{8}$$

$i = 1, 2, 3, 4$.

The final ranks are assigned taking into account the values of average relative distances. The institution with the lowest average relative distance receives the rank 1. The location of each institution compared to the one with the best performance is a ratio: the average relative distance over the lowest average relative distance.

$$loc_i^{\%} = \frac{\bar{d}_i}{\min(d_i)_{i=1,4}} \cdot 100 \tag{9}$$

Wilcoxon Signed Ranks test and Kruskal-Wallis test are nonparametric tests used when the series repartition is not known or non-normal. These tests are applied to check the differences between populations. In this case, the differences between the real values and the forecasted ones are checked using the two non-parametric tests. The null hypothesis refers to the lack of differences, while the alternative one shows that there are significant differences between the forecasts and the registered values. A p-value that is lower than 0.05 implies the rejection of null hypothesis. For small samples the chi-square approximation gives better results in most cases than Kruskal-Wallis test, according to Conover [6].

Comparisons between forecasts can be done using binary logistic regression when the dependent variable is a qualitative one. For this type of regression some assumptions are not considered (errors non-correlation, normality or homoscedasticity).Odd-ratios (OR) are computed to see how much the occurrence chances of an alternative of the dependent variable modify when the independent variable change with one unit. The coefficient of the exogenous variable from the regression model is denoted by b_1 .

If OR is higher than 1, an increase by one unit in the level of the exogenous variable implies a growth by $e^{\hat{b}_1}$ in the level of the dependent variable.

The absolute errors for each year in the forecasting horizon are computed. We test if each error differs significantly from a threshold fixed at 0.5%. We choose a threshold of 1%. The dependent variable has two alternatives:

$$\text{error_significance} = \begin{cases} 1 & \text{high error} \\ 0 & \text{low error} \end{cases}$$

For each forecasted value of the variable, the significance of the error is computed.

The Spearman's and Kendall's coefficients of correlation might be computed to see the associations between the real values and the predicted ones.

3. Inflation Forecasts Comparisons for USA

The forecasting horizon for USA inflation rate is 1978-2012 and the predictions are provided by Survey of Professional Forecasters (SPF), Congressional Budget Official (CBO), Blue Chips (BC) and Administration.

It is not recommended the use of a single measure of accuracy. In our study 5 accuracy indicators were selected.

Table 1. The accuracy of inflation forecasts provided by SPF, CBO, BC and Administration (1978-2012)

Accuracy measure	SPF	CBO	BC	Administration
ME	2.9826	0.123	0.4527	0.6878
MAE	3.0597	1.8365	1.4673	1.4376
RMSE	4.1058	2.6445	2.5094	2.5279
U1	0.4525	0.3053	0.2985	0.2591
RRSSE	0.3467	0.3178	0.2856	0.2467
U2	0.8084	73.3077	152.8537	192.5686

According to U1 statistic, the best forecasts are provided by Administration, being followed by BC anticipations, CBO ones and finally the SPF forecasts. However, CBO predictions have the lowest mean error. Administration registered the lowest mean absolute error for their predictions. Only the SPF anticipations are better than the naïve forecasts. The indicator RRSSE introduced by us in literature gave the same results as U1.

The multi-criteria ranking solves the problem of contradictory measures of accuracy by considering their influence at the same time.

Table 2. The ranks method for the comparison of USA inflation forecasts accuracy (1976-2012)

Criterion	SPF ranks	CBO ranks	BC ranks	Administration ranks
ME	4	1	2	3
MAE	4	3	2	1
RMSE	4	3	1	2
U1	4	3	1	2
U2	1	2	3	4
Sum of ranks	17	12	9	12
Final rank	4	2	1	3

The ranks method recommends BC forecasts as the best and the SPF as the less accurate. CBO and Administration predictions have the same degree of accuracy.

According to the second method of multi-criteria ranking, the best forecasts on the horizon 1976-2012 were

provided by CBO. The hierarchy of institutions is continued by: SPF, BC and Administration. So, there are differences regarding the hierarchy provided by the two methods. In general, the method of relative distance according to the best institution gives better results. However, CBO gave the best performance according to both methods.

Table 3. The method of relative distance related to the maximal performance for the comparison of USA inflation forecasts accuracy (1976-2012)

Criterion	SPF ranks	CBO ranks	BC ranks	Administration ranks
ME	24.2542	1.0000	3.6813	5.5934
MAE	2.1284	1.2775	1.0207	1.0000
RMSE	1.6362	1.0539	1.0000	1.0074
U1	1.7463	1.1783	1.1519	1.0000
U2	1.0000	90.6777	18839.741	2.39E+09
Average relative distance	2.7149	2.7013	9.600	106.0963
Final rank	2	1	3	4
Location (%)	1.0050	1.0000	3.5538	39.2753

The dependencies between the effective values and the forecasted ones are analyzed using non-parametric tests like Wilcoxon Sum Ranks and Kruskal-Wallis.

After the application of non-parametric tests we made the following conclusions with a probability of 95%:

The differences between CBO predictions and the registered values are not significant;

The differences between SPF predictions and the real values are not significant, but the Significance indicator is lower; this shows that CBO predictions are better than the SPF ones;

The differences are not significant between BC forecasts and the real values, but the p-value is lower than that of the other two predictions; this implies that SPF and CBO expectations are better than BC ones;

There are significant differences between Administration forecasts and the effective values on inflation. The results of the tests applied in SAS are presented in **Appendix 1**.

So, the hierarchy given by the application of non-parametric tests is: Administration, CBO and Blue Chips. For all the SPF predictions the errors are significant and larger than the threshold of 1%.

The odds of having a low error for CBO forecasts grow with 36.6%, while the chances for Blue Chips increase with 25.3%. For Administration forecasts only few errors are not significant. So, the hierarchy provided by the analysis of binary logistic regression is: CBO, Blue Chips, Administration and SPF. The results of this procedure are displayed in **Appendix 2**.

The Spearman's and Kendall's coefficients of correlation are computed in **Appendix 3**. The strong correlation is between BC forecasts and real value, being followed by the one between Administration and the effective values and CBO and real values. The correlation between SPF expectation and the real values is not significant.

4. Strategies of Improving the Forecasts Accuracy

Bratu (2012) specified some strategies of improving the forecasts accuracy (application of filters and exponential smoothing techniques, combined forecasts, regressions models, historical errors method).

The most use approaches for combined forecasts are: optimal combination (OPT), inverse MSE weighting scheme (INV) and equal-weights-scheme (EW).

Bates and Granger) used two forecasts $f_{1,t}$ and $f_{2,t}$, of the same variable X_t , derived h periods ago. For unbiased forecasts, the error is computed as: $e_{i,t} = X_{i,t} - f_{i,t}$.

The errors follow a normal repartition of parameters 0 and σ_i^2 . If ρ is the errors coefficient of correlation the covariance is $\sigma_{12} = \rho \cdot \sigma_1 \cdot \sigma_2$. The linear combination of the two predictions is: $c_t = m \cdot f_{1t} + (1-m) \cdot f_{2t}$.

The error of the combined forecast is: $e_{c,t} = m \cdot e_{1t} + (1-m) \cdot e_{2t}$. The combined forecast mean is zero and the variance is:

$$\sigma_c^2 = m^2 \cdot \sigma_1^2 + (1-m)^2 \cdot \sigma_2^2 + 2 \cdot m \cdot (1-m) \cdot \sigma_{12}$$

The optimal value for m is determined by minimizing the error variance (m_{opt}) [2]:

$$m_{opt} = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2 \cdot \sigma_{12}} \tag{10}$$

The inverse weight (m_{inv}) is computed as:

$$m_{inv} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{11}$$

For equally weighted combined predictions (EW) the same weights are given to all models.

The U1 Theil's coefficient is computed for the combined forecasts based on the three schemes.

Table 4. The accuracy of USA inflation combined forecasts on the horizon 1976-2012

Combined forecasts	U1(OPT scheme)	U1 (INV scheme)	U1 (EW scheme)
SPF+CBO	0.6035	0.6219	0.3888
SPF+BC	0.6055	0.6139	0.4022
SPF+ADMINISTRATION	0.6299	0.6583	0.4318
CBO+BC	0.2493	0.2883	0.2890
CBO+ADMINISTRATION	0.2983	0.2999	0.3003
BC+ADMINISTRATION	0.3200	0.3011	0.3009

The combined forecasts of CBO and BC using OPT scheme improved the accuracy of all predictions. This type of combination gave better results than all the initial forecasts, excepting Administration ones. All the other combined forecasts excepting those where SPF

anticipations are implied are better than SPF and CBO ones.

The application of filters to the initial forecasts and also the exponential smoothing techniques as Holts Winters are utilized for improving the forecasts accuracy [4].

The Hodrick–Prescott (HP) filter is used to extract the trend of the data series. Razzak explained that the Hodrick-Prescott filter is a true 'filter' at the end of the sample and a 'smoother' one over the sample [10]. The output gap from the true filter determines more accurate out-of-sample predictions of inflation. Christiano and Fitzgerald showed that Band-Pass filter is used to determine the component of the time series that is situated within a specific band of frequencies [5]. Christiano-Fitzgerald filter (CF filter) converges on long run to an optimal filter. It has a steep frequency response function at the band boundaries.

Holt-Winters (HW) Simple exponential smoothing technique is recommended for data set with linear trend and no seasonal variations. The filters and the HW technique are utilized to smooth the predictions provided by the four institutions. Then, the accuracy of the new forecasts is evaluated.

Table 5. The U1 values of USA inflation forecasts on the horizon 1976-2012

Technique of smoothing	SPF	CBO	BC	Administration
Hodrick-Prescott filter	0.6859	0.2474	0.2604	0.2784
Christiano-Fitzgerald filter	0.9656	0.9817	0.9639	0.9687
Baxter King filter	0.8976	0.8649	0.8535	0.8569
Holt-Winters technique	0.6651	0.2988	0.2878	0.3010

The Hodrick-Prescott technique and the Holt-Winters model improved the accuracy of BC and CBO forecasts. The great improvement was generated by Hodrick-Prescott filter for both types of predictions. For CBO anticipations, the accuracy is even better than that of the predictions provided by the other institutions or by the combined forecasts.

5. Conclusions

This research enriches the literature regarding the assessment and the improvement of forecasts accuracy.

According to U1 statistic and the new introduced RRSSE indicator and according to ranks method and to Spearman's coefficient of correlation the hierarchy of institutions that forecasted between the two-year inflation in 1982-2011 is: CBO, Administration and Blue Chips. The relative distance method with respect to the better institution, the logistic regression, the non-parametric tests provided the following ranking: Administration, CBO and Blue Chips. The highest improvement in accuracy was brought by the combined forecasts of Blue Chips and Administration using inverse MSE scheme. The smoothed predicted values based on Holt-Winters technique, Hodrick-Prescott, Baxter King and Christiano-Fitzgerald filters did not improve the forecasts accuracy.

The novelty of this research consists in the application of some statistical approaches to compare the predictions accuracy, these methods never being mentioned in

literature in this context. The results of the new approach are better than those provided by the U Theil's statistic, because more aspects of accuracy problem are taken into account.

References

- [1] Abreu I., "International organizations' vs. private analysts' forecasts: an Evaluation", *Banco de Portugal Papers*, 29-34, 2011.
- [2] Bates, J., and C. W. J. Granger, "The Combination of Forecasts", *Operations Research Quarterly*, 20(4): 451-468, Jul. 1969.
- [3] Bratu, M. (2012). *Strategies to Improve the Accuracy of Macroeconomic Forecasts in USA*, LAP LAMBERT Academic Publishing, 2012, 6-28.
- [4] Bratu (Simionescu) M., "Filters or Holt Winters technique to improve the forecasts for USA inflation rate ?", *Acta Universitatis Danubius. Economica*, 9(1): 23-45, 2013.
- [5] Christiano, L. J. and Fitzgerald, T.J., "The Band Pass Filter", *International Economic Review*, 44(2): 435-465, 2003.
- [6] Conovor W.J., *Practical nonparametric statistics*, Wiley series in probability and statistics, New York: Wiley, 1999, 56-78.
- [7] Edge R.M., Kiley M.T. and Laforte J.-P., "A comparison of forecast performance between Federal Reserve Staff Forecasts simple reduced-form models and a DSGE model", *Finance and Economics Discussion Series*, 85-89, 2009.
- [8] Hodrick R. and Prescott, E.C., Postwar U.S. Business Cycles: An empirical investigation, *Journal of Money, Credit and Banking*, 1(16): 84-90, 2003.
- [9] Österholm, P., "The limited usefulness of macroeconomic Bayesian VARs when estimating the probability of a US recession", *Journal of Macroeconomics*, Elsevier, vol. 34(1): 76-86, 2012.
- [10] Razzak W., The Hodrick-Prescott technique: A smoother versus a filter: An application to New Zealand GDP, *Economics Letters*, 57(2), 163-168, 1997.
- [11] <http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/115xx/doc11553/forecastingaccuracy.pdf>.

Appendix 1

The results of non-parametric tests in SAS

Institution	Chi-Square	Asymptotic Pr> Chi-Square	Exact Pr>= Chi-Square
CBO	24.112	0.197	0.218
SPF	22.336	0.185	0.203
BC	21.0563	0.177	0.149
Administration	1.234	0.027	0.033

Appendix 2

Binary logistic regressions in SPSS

		Sig.	Exp(B)
Step 1 ^a	cbo	.002	.366
	Constant	.001	205.153

a. Variable(s) entered on step 1: cbo.

		Sig.	Exp(B)
Step 1 ^a	Bc	.007	.253
	Constant	.002	510.315

a. Variable(s) entered on step 1: bc.

Appendix 3

Spearman's and Kendall's coefficients between the real values and the forecasts of the four institutions

Correlation between real values and the forecasts of:	Spearman's coefficient	Kendall's coefficient
CBO	0.404 (0.013)	0.263 (0.025)
BC	0.658 (0.000)	0.465 (0.000)
SPF	-0.117 (0.491)	-0.082 (0.48)
Administration	0.656 (0.000)	0.452 (0.000)

The Sig. (2-tailed) are in brackets