Evaluation of GDP Growth Forecasts: Does Using Different Data Vintages Matter?[1]

*Jiří ŠINDELÁŘ – Petr BUDINSKÝ**

Abstract

This paper deals with effect of different real-time data vintages (recent and first outturn) on accuracy of real GDP growth forecasts produced by main Czech and Slovak public authorities (ministries of finance, central banks). Firstly, variation in the real-time data itself was analysed, along with of multidimensional forecasting error evaluation (MAE, RMSE, MASE measures). Then, battery of statistical tests was applied in order to determine, whether the switch from first to recent real-time data affects forecasts´ accuracy in a significant manner (Wilcoxon Signed Rank test, Sign test) and whether it affects relative accuracy between individual institutions (Kruskal-Wallis test, Mann-Whitney U test). Our results show that while the change in underlying data affects forecasting accuracy in our sample (using recent data lead to higher errors), the changes were neither found statistically significant in strong majority of surveyed cases, nor affected the relative accuracy of involved institutions.

Keywords: *GDP, forecasting, forecast evaluation, data vintages*

JEL Classification: E27, C53, E01, E62

Introduction

Keeping a track record of forecasting accuracy is an important part of the forecasting process. This is even more so with macroeconomic predictions produced by public institutions, which play a crucial role in every country´s fiscal and monetary policies, affecting the business sector as well. Evaluating macroeconomic predictions, however, contains serious challenges by design in terms of underlying data choice. Inevitably, every researcher faces a decision whether to use the most recent, actual data (i.e. recent out-turn), or utilise historical, real-time

* Jiří ŠINDELÁŘ – Petr BUDINSKÝ, University of Finance and Administration Prague, Estonská 500, 101 00 Prague 10, Czech Republic; e-mail: jiri.sindelar@vsfs.cz; petr.budinsky@vsfs.cz

[1] This article was created with the contribution of institutional support for long-term conceptual development of the research organization University of Finance and Administration.

data available at a given period of the forecasting horizon (i.e. first out-turn). Both options have potential interpretational and methodological trade-offs, offering advantages and disadvantages from the analytical point of view, and a potentially significant impact on results and their usage. Furthermore, the above mentioned is emphasized by business cycles and macroeconomic policy impacts, as well as improvements made to the fundamental forecasting models, both of which occur between periodic data revisions.

This problem is particularly important in relation to Gross domestic product (GDP) growth forecasting, where a number of different approaches can be observed in prolific papers dealing with forecasts´ evaluation (e.g. Öller and Barott, 2000; Allan, 2013; Daníelsson, 2008 and others). As assumed by most authors, different data vintage can significantly affect the outcome of the evaluation (Mc Nees and Ries, 1983; Kenen and Schwarz, 1986 or Artis, 1996). Unfortunately, there is limited empirical evidence available, with most of the relevant papers offering only subjective or very general outcomes. In the Central-Eastern European context, the specific evidence is missing altogether, with authors relying on actual (Antal, Hlaváček and Horvath, 2008; Antoničová et al., 2009; Feldkircher et al., 2015) or real-time (Arnoštová et al., 2011) data without almost any reasoning, more or less on an intuitive basis.[2] This, coupled with the relatively outdated character of principal publications listed above, creates an important research opportunity.

Therefore, this paper seeks to provide empirical evidence on the effect that different data inputs (recent vs. first out-turn data) have on GDP growth forecast accuracy, in the context of Czech and Slovak GDP forecasts. The paper is divided into three sections. In the first part, an overview of relevant literature is provided, highlighting the main theoretical approaches and deriving main study assumptions. In the second part, a battery of forecasting errors (AFE, RMSE, MASE[3]) is computed and statistical analysis of differences between both methodological concepts is undertaken (Kruskal-Wallis test, Wilcoxon Signed Rank test). Finally, the outcomes of the analysis are discussed with regard to the theoretical assumptions and recommendations for further research are provided.

Literature Overview

As mentioned above, although the decision on whether to use recent or historic real variable data bears important consequences, problem coverage in the concurrent literature is rather sparse. As summarized by Table 1, six empirical studies spanning almost over the last 20 years have been identified:

[2] Rusnák´s (2016) paper being a rare exception.

[3] AFE – Average Forecasting Error; RMSE – Root Mean Square Error; MASE – Mean Absolute Scaled Error.

Table 1

Concurrent Studies Meta-analysis

Study	Method	Surveyed sample	Results
Robertson and Tallman (1998)	RMSE, correlation analysis (real time vs. final data, VAR/AR/Advanced estimate model setup)	Real U.S. GDP growth forecasts, one and two quarters ahead (1959 – 1977, 1977 – 1998)	(1) Use of the latest vintage data does not cause the model forecast (VAR) to be much different. (2) The choice of the data does alter the measured forecasting accuracy, but not too seriously – it does not change the model ranking: • $\Delta_{RMSE/VAR} = –0.46/–0.4^{a}$ • $\Delta_{RMSE/AR} = –0.52/–0.53^{a}$ • $\Delta_{RMSE/AE} = –1.15^{a}$; $\Delta_{CORRELATION/AE} = –0.13^{b}$
Croushore and Stark (2001)	RMSE (real time vs. final data, simple ARIMA model setup)	Real U.S. GDP growth forecasts, one year ahead (1958 – 1999)	The forecasting error is not very different, when forecasts are based on real-time data as opposed to final revised data. The model based on final data actually produced a slightly less accurate forecast than the real-time one: • $\Delta_{RMSE/ARIMA} = 0.02$
Aruoba (2008)	RMSE, Diebold--Mariano test (real time setup, moving average/naïve-zero models for revisions forecasting)	Real U.S. GDP growth forecasts, one quarter and one year ahead (1958 – 1999)	The study finds that even the simple model (moving average) is able to predict revisions in data better than the naïve-zero forecast, but the difference is marginal and not statistically significant (real GDP growth): • $\Delta_{RMSE/MA\text{-}NZ} = –0.04$ (annual value, not statistically significant difference – DM_t) • $\Delta_{RMSE/MA\text{-}NZ} = –0.02$ (quaterly value, not statistically significant difference – DM_t)
Diron (2008)	RMSE (real time vs. final data setup, multiple models)	Real Euroarea GDP growth forecasts, one quarter ahead (2001 – 2004)	(1) Forecasts for individual quarters tend to be similar whether they are based on preliminary or revised data. (2) Revisions to the monthly variables and to GDP growth account for only a small share of the overall forecast errors: • $\Delta_{RMSE/AR} = –0.05$ • $\Delta_{RMSE/AverageForecast} = –0.06$
Taylor (2014)	RMSE, t-test, Monte Carlo simulation (real time vs. final data setup, four models)	Real U.S. GDP growth forecasts, one and two quarters ahead (1959 – 1977, 1977 – 1998)	The results indicate the model performance is somewhat sensitive to data vintage noise (extent of revisions). In three out of twelve comparisons, the most accurate model (VAR based on ADS index) did not confirm its superior forecasting efficiency in terms of RMSE differential, when fed with the vintage data: • low noise: significant superiority of VAR_{ADS} confirmed in 4/4 comparisons (three times at $p = 0.01$; single at $p = 0.05$) • medium noise: significant superiority of VAR_{ADS} confirmed in 3/4 comparisons (twice at $p = 0.01$; single at $p = 0.05$) • high noise: significant superiority of VAR_{ADS} confirmed in 2/4 comparisons (twice at $p = 0.01$)
Raponi and Frane (2014)	RMSE, Diebold--Mariano test (real time vs. final data setup, four models)	Real Italian GDP growth forecasts, 1 – 3 months ahead (2006 – 2013)	The results indicate that the forecasting model based on revised data (AR_{REV}) exhibited significantly better performance in 1-step (month) ahead forecast, with mixed results in two and three step ahead forecasts: • $\overline{\Delta}_{RMSE/1\text{-}step} = –0.3$; $p_{DM} = 0.001 – 0.045^{c}$ • $\overline{\Delta}_{RMSE/2\text{-}step} = 0.03$; $p_{DM} = 0.352 – 0.926^{c}$ • $\overline{\Delta}_{RMSE/3\text{-}step} = –0.12$; $p_{DM} = 0.284 – 0.901^{c}$

[a] The Δ value is calculated as a difference between recent out-turn (RO) and first out-turn (FO) results (one quarter / two quarters ahead forecast).

[b] The difference of the correlation between forecast and real value (RO – FO). Just one quarter results available for the AE model.

[d] Range of Diebold-Mariano test probabilities, when comparing AR_{REV} with the other three models.

VAR – Vector autoregressive model
AR – Autoregressive model
ARIMA – Autoregressive integrated moving average model
AE – Advanced estimated model
MA – Moving average model

NZ – Naïve-zero model
RMSE – Root mean squared forecasting error
ADS – Philadelphia Fed Aruoba-Diebold-Scotti index.
DM – Diebold-Mariano test

Source: Own research.

Going through the highlighted papers, several common patterns are evident. Firstly, the majority of available research is focusing mainly on forecasting models´ performance under different data setup (i.e. investigating differences in accuracy, when using recent out-turn data as an input, so called now-casting), instead of analysing the effect of different data vintages on accuracy of existing forecasts. This is a vocal point of concurrent forecasting research in this context, as evinced by other papers not listed above (e.g. Croushore and Stark, 2003; Feng, 2005 or Giannone, Reichlin and Small, 2008). Secondly, nearly all of the available papers deal with short horizon forecasts only, essentially one to two quarters ahead being the average maximum. Geographically, all of the papers listed above are centred in Western Europe and the U.S., with two of the U.S. papers (Croushore and Stark, 2001; Aruoba, 2008) explicitly sharing a similar data set.[4]

From a factual perspective, the results mostly indicate only minor differences among themselves when using different vintages of the data. This creates an interesting schism with older theoretical works, which generally presume significant effect, in terms of accuracy differences[5] (e.g. McNees and Ries, 1983; Kenen and Schwarz, 1986). Combining both inputs together, we find that most papers recommend using historical data "available at the time of the forecast"; the reasoning behind operates mainly with the forecaster´s aim to hit the first value and not the latter one, incorporating unforeseen revisions (Artis, 1996). Related applied studies generally oscillate between preliminary estimates (usually published before the end of the forecast period – see e.g. Di Fonzo, 2005; McKenzie and Gamba, 2008) and first available/settled estimates (usually published during the first/second half of the following period – e.g. Keereman, 1999; Öller and Barott, 2000 or Antal, Hlaváček and Horvath, 2008). The advocates of the opposite approach, while less numerous (e.g. Daníelsson, 2008 or Arnoštová et al., 2011), simply rely on the latest published data available, without further restrictions. They argue that "it is crucial to use the most accurate estimate of the actual data in order to avoid penalising the best prediction of what actually happened as opposed to the best prediction of what initially was mistakenly thought to have happened" (McNees and Ries, 1983, p. 27).

Finally, methodological aspects of the highlighted studies point to a common evaluation algorithm. All of the papers rely on Root Mean Squared Error (RMSE), in some cases accompanied by the prolific Diebold-Mariano (DM) test.

[4] Real time data set, published by the Federal Reserve Bank of Philadelphia in 1999 – see Croushore and Stark (2001) for details.

[5] This expectation is underlined by the extent of real GDP growth revisions, which is commonly found (Aruoba, 2008) to reach as much as 0.25% mean and 1.51% std. deviation value, and to be statistically significant (Q-statistics at 20 lags, $p = 0.00$).

Both of the modules are problematic in a certain sense. While using RMSE as the main metrics is not necessarily a bad strategy (as it well covers error magnitude in discontinuities), it offers just a single perspective on the final accuracy when the multi-metrical approach is widely recommended (Makridakis and Hibon, 1995; Armstrong, 2001). The application of the DM test brings up even more serious issues, as the method even in its small sample variant (Harvey, Leybourne and Newbold, 1997), exhibits serious issues when confronted with finite time-series and serial persistence (Christensen et al., 2007). Both are likely to occur in GDP growth timelines. Because of this, the evaluation framework found in most studies can be substantially enhanced, either by using more measures from the scale-dependent or relative family, and by using more suitable, non-parametric test methods.

In the Czech-Slovak context, the problem is generally not covered, with the single exception of Rusnák´s (2016) paper. In this study, a comparison of two models´ performance (dynamic factor model, judgmental now-cast and their combination) on the basis of multiple indicators led by real GDP growth is undertaken (2005 – 2012 period). By using multiple horizons (1Q – 6Q ahead), Rusnák came to similar results like the studies presented earlier, with Δ_{RMSE} oscillating around 0.1 – 0.3 difference in favour of the models based on final data, slightly increasing with the lengthening horizon. None of these were found statistically significant, using the problematic Diebold-Mariano test.

Overall, both recent and historical data approaches exhibit analytical merit and have found supporters among the scientific community. Most of the empirical papers, however, have diverted their focus to the performance of different models with different data inputs, which, although a vital topic itself, does not directly answer the question of whether and how the different data benchmarks affect the accuracy of individual forecasts. Also, most papers share methodological deficiencies related to the measures used, which might have affected their final outcome. We next present relevant hypotheses derived from the presented research and set of methods utilised in our following analysis.

Research Hypotheses

Based on the available research, in the first place, we presume that there will be no statistical difference between absolute (MAE) and relative (MASE) accuracy measures, when using different vintages of the real value (Y_t) data:

H_1: *Forecasting accuracy measured by MAE metrics is not significantly different when using first and recent out-turn Y_t data.*

H_2: *Forecasting accuracy measured by MASE metrics is not significantly different when using first and recent out-turn* Y_t *data.*

Furthermore, we will examine the differences between the surveyed institutions themselves, both in the real-time and actual data setup:

H_3: *Forecasting accuracy measured by MAE metrics is not significantly different among surveyed institutions, when using first out-turn* Y_t *data.*
H_4: *Forecasting accuracy measured by MASE metrics is not significantly different among surveyed institutions, when using first out-turn* Y_t *data.*
H_5: *Forecasting accuracy measured by MAE metrics is not significantly different among surveyed institutions, when using recent out-turn* Y_t *data.*
H_6: *Forecasting accuracy measured by MASE metrics is not significantly different among surveyed institutions, when using recent out-turn* Y_t *data.*

While the first two hypotheses form the bulk of response in relation to our main research question, the additional four indirectly augment the findings by analysing causal difference between individual institutions. This also adds another layer of factual findings itself. The data sources and methods employed in the verification are presented next.

Data

In our analysis, we use annual real GDP growth forecasts produced by four Czech and Slovak institutions, in the following periods:

- Ministry of Finance, Czech Rep. (MFCR, 2016) – 1995 – 2014 period
- Czech National Bank (CNB, 2016) – 1998 – 2015 period
- Ministry of Finance, Slovak Rep. (MFSK, 2016) – 2004 – 2014 period
- National Bank of Slovakia (NBS, 2016) – 1995 – 2014 period.

From the variety of forecasts, the institutions publish each year (usually four forecasts with multiple horizons), the paper utilizes the autumn forecast, published mostly in October/November. Because we use annual, next year forecasts, the surveyed forecasting horizon is determined within 15 months (15M). Such a setup not only covers the most important yearly prediction used for the budgetary and other fiscal procedures, it also follows the practice of prolific papers cited, such as Öller and Barott (2000) or Keereman (1999).

Regarding the crucial real value indicator, we utilise two variants of data:

- ***Most recent out-turn*** – most recent GDP growth data provided by the Czech Statistical Office (CZSO, 2016) and the Statistical Office of the Slovak Republic (SOSR, 2016) respectively.[6]

[6] Data available by 3/2016.

- ***First out-turn*** – historical GDP growth data in terms of a preliminary (first available) estimate, published by the end of next year (i.e. the year being forecast, usually in December), provided by the OECD (2016) Economic Outlook.

Similarly to the previous, by adopting the first estimate strategy, the study puts itself on a comparative basis with vital parts of the papers forming current theoretical background (aforementioned Di Fonzo, 2005; McKenzie and Gamba, 2008 and others). From a practical perspective, the first GDP estimate is also a value that many institutions measure their efforts with and represents the most contrasting preliminary figures available. This further underlines the comparative framework used in statistical testing.

Method

In terms of accuracy evaluation, the study method relies on a combination of three forecasting error measures, from the scale-dependent and relative family respectively. Denoting the real value at a given time as Y_t, forecast value as F_t and the subsequent forecasting error as $E_t = (Y_t – F_t)$, we can define them as the following:

- ***Mean Absolute Error (MAE)***

$$MAE = mean(|E_t|)$$

- ***Root Mean Squared Error (RMSE)***

$$RMSE = \sqrt{mean(E_t^{\,2})}$$

- ***Mean Average Scaled Error (MASE)***

$$MASE = mean \frac{E_t}{\frac{1}{n-1}\sum_{i=2}^{n}|Y_i - Y_{i-1}|}$$

These measures form the backbone of “raw” forecast accuracy evaluation, reflecting individual weaknesses and recommendations presented in summary papers such as Hyndman and Koehler (2006) and Armstrong and Collopy (1992). The most relevant limitation includes forecast and real values close to or equal to zero, which raises a natural restriction to the usability of some measures, namely from the percentage error family. For this reason, the evaluation benchmark, proposed in literature (Fair, 1986; Fildes and Stekler, 2000; Öller and Barot, 2000), relies on a combination of squared scale-dependent measures (RMSE, imposing heavier penalties on bigger errors, thus ensuring higher contrast), absolute measures (MAE, to detect possible over-forecasting and sandbagging) and relative measures (MASE, to simply judge the forecaster´s ability to perform better than the benchmark, usually the naïve method). Such a combination covers all of the basic accuracy features, with the gap arisen by the exclusion of percentage errors usually covered by other measurement supplements and/or relative measures.

The second part of the analysis is composed of statistical verification of differences between individual measures, and their significance, in relation to set hypotheses. A battery of two non-parametric methods is utilised:

(i) For evaluation of differences between error measures calculated with the most recent and first out-turn data (related samples), we utilise the ***Wilcoxon sign test***. Working with its classical variant, as defined in Wilcoxon (1945), this test indicates, whether indicators originating from the same sample under diverse conditions differ (Alternative hypothesis – diff. of medians not equal to zero), or not (Null hypothesis – diff. of medians equal to zero). Before applying the test itself, we have also verified its prerequisite in terms of data symmetry, by applying the Miao, Gel and Gastwirth test (Miao, Gel and Gaswirth, 2006), with positive results: the symmetry was not rejected for every included time series (i.e. H_0 was not rejected on common p = 0.05; full disclosure of the p-values can be found in Appendix 1). Finally, we have also used the simple Sign test fulfilling the same role as the Wilcoxon sign test, to act as our control method.

(ii) Regarding differences between individual institutions´ forecast accuracy, the ***Kruskal-Wallis test*** was utilised, enabling us to compare accuracy of forecasts between institutions. The decision to omit the more renowned Diebold-Mariano method was based on aforementioned evidence of its deficiencies (rejecting null too often – oversized type I error), when dealing with short samples containing serial persistence (Christensen et al., 2007). For detailed analysis of differences between individual institutions, the two-tailed Mann-Whitney U test (Mann and Whitney, 1947) was applied. In order to verify both tests´ prerequisites, we further evaluated independence of the time series (institutions) using the Ljung-Box independence test (Ljung and Box, 1978). Again, a positive result was obtained (i.e. H_0 was not rejected on common p = 0.05; full disclosure of the p-values can be found in Appendix No. 1), combined with data symmetry tested earlier.

All of the tests were conducted separately for the MAE and MASE measures; since RMSE is a naturally interval metrics, it was omitted from this part of the analysis. We next present results gained in both steps.

Results

Data Vintages Effect on Underlying Data

As aforementioned, some studies point to a substantial range of changes the overtime revisions can inflict on the first out-turn data. In case of real GDP growth in the Czech Rep. (CZ) and Slovakia (SK), the effect was quite heterogeneous and surprisingly, not predominantly related to the older periods. The following table summarizes basic descriptive statistics (mean, standard deviation) of both absolutized (ABS) and non-absolutized (NONABS) differences between first and recent out-turn data (Table 2):

Table 2

Differences between First and Recent Data Vintages (% of GDP, underlying Y_t data)

Period	Czech Republic				Slovak Republic			
	Mean (NONABS)	Mean (ABS)	StDev (NONABS)	StDev (ABS)	Mean (NONABS)	Mean (ABS)	StDev (NONABS)	StDev (ABS)
1995 – 1999	0.482	1.319	1.431	0.736	0.556	0.660	1.193	0.667
2000 – 2004	0.628	0.970	0.926	0.559	0.405	0.681	0.710	0.452
2005 – 2009	–0.077	1.004	1.140	0.545	0.316	0.975	1.091	0.583
2010 – 2015	0.035	0.289	0.439	0.332	0.116	0.554	0.662	0.380
1995 – 2015	**0.256**	**0.866**	**1.068**	**0.676**	**0.132**	**0.799**	**0.969**	**0.564**

Source: Own calculations.

Coming from the numbers at hand, we can derive several important findings. Firstly, the prevailing direction of revisions is the same in both countries: in the CZ the overall effect (mean of non-absolutized values) led to an increase compared to the first out-turn data, in SK the outcome was the same and the first estimates were aggregately toned up as well, albeit not as much. Regarding the magnitude of the revisions (mean of absolutized values) in both countries, the general size of revisions points to slightly less than one percentage point. This is a rather substantial change, keeping in mind that growth values generally oscillate between zero and five percent. Finally, coming back to the prelude, the distribution of revisions was not tailing towards older values, which is particularly true for Slovakia, when the extent of the revisions in 2005 – 2009 greatly surpassed the first two periods. Still, we can conclude that the overall extent of revisions in both countries is considerable, also because of high variation indicated by the standard deviation values. How will it affect the institutions´ forecasting performance?

Accuracy Measures

From the accuracy perspective, the forecasts compared with the first (FO) and the recent (RO) out-turn data offer visible differences, as outlined in Table 3. The calculation of accuracy metricises was performed for every year of the series as well as for a period of five years, tied to the initial year 1995. With the exception of RMSE, which, as a naturally interval measure, was calculated only in the five year term.

Table 3

Accuracy Measures Results – Czech Republic

Year	**MFCR**						**CNB**					
	AE_{RO}	AE_{FO}	$RMSE_{RO}$	$RMSE_{FO}$	$MASE_{RO}$	$MASE_{FO}$	AE_{RO}	AE_{FO}	$RMSE_{RO}$	$RMSE_{FO}$	$MASE_{RO}$	$MASE_{FO}$
1995	0.5	0.0			0.3	0.0						
1996	5.7	4.1			1.1	1.1						
1997	2.5	2.9			7.0	1.8						
1998	0.4	2.3			0.2	11.5	1.4	0.5			0.8	2.5
1999	2.9	1.1			1.0	0.4	3.6	1.8			1.3	0.6
1995 – 1999	2.4	2.1	3.1	2.5	1.9	2.9	2.5	1.2	2.7	1.3	1.0	1.6
2000	0.1	0.0			0.0	0.0	0.6	0.5			0.4	1.0
2001	2.2	1.3			1.5	2.6	2.3	1.4			1.6	2.8
2002	0.3	0.8			0.2	*	1.9	0.8			1.0	*
2003	2.1	1.1			1.6	0.8	2.1	1.1			1.6	0.8
2004	2.8	1.2			1.9	1.3	2.1	0.4			1.4	0.5
2000 – 2004	1.5	0.9	1.9	1.0	1.0	1.2	1.8	0.8	1.9	0.9	1.2	1.3
2005	2.5	1.8			5.7	1.3	3.1	2.4			7.1	1.7
2006	0.6	1.2			0.5	12.0	0.0	0.6			0.0	6.0
2007	2.3	0.6			0.8	0.4	2.3	0.6			0.8	0.4
2008	8.5	8.1			1.1	0.9	7.7	7.3			1.0	0.8
2009	2.0	2.1			0.3	0.3	0.9	1.0			0.1	0.1
2005 – 2009	3.2	2.8	4.2	3.9	1.7	3.0	2.8	2.4	3.9	3.5	1.8	1.8
2010	0.0	0.1			0.1	0.3	0.8	0.9			2.3	3.0
2011	1.9	1.9			0.7	0.6	2.1	2.1			0.7	0.7
2012	1.2	2.2			3.3	3.7	0.7	1.7			2.0	2.8
2013	0.7	1.1			0.3	0.3	0.1	0.3			0.0	0.1
2014	1.7	1.8			0.8	0.9	1.7	1.8			0.8	0.9
2010 – 2014	1.1	1.4	1.3	1.6	1.0	1.2	1.1	1.4	1.3	1.5	1.2	1.5
1995 – 2014	2.0	1.8	2.8	2.5	1.4	2.1	2.0	1.5	2.4	2.0	1.4	1.5

* Because of data attributes, MASE cannot be calculated for those years.

Source: Own calculations.

Both Czech institutions exhibit interesting summary data. Firstly, the CNB was able to reach lower scale-dependent forecasting errors (MAE, RMSE) in almost every surveyed period. MFCR, on the other hand, achieved a slightly better performance vs. the naïve benchmark (MASE) in some sub-periods, particularly with the first outturn data. Regarding the crucial difference between RO and FO accuracy, results indicate two important findings: (i) the differences were rather minor in scale (mostly oscillating around one point of error) and (ii) with the exception of the last sub-period (2010 – 2014), the forecasting error always increased, when switching from FO to RO input data. From an economic point of view, forecasting errors, understandably, increased in uncertain macroeconomic periods, which are predominantly concentrated around the 1996 – 1997 recession, the 2008 – 2009 global financial crisis and subsequent fiscal consolidation and the 2013 CNB exchange rate intervention. Finally, we cannot conclude that the differences between paired error measures decrease over time, i.e. are higher with older data. Although with the CNB such a pattern is observed, MF exhibits different results, error differences in later periods (2000 – 2004; 2005 – 2009) being higher than the older ones.

Table 4

Accuracy Measures Results – Slovak Republic

Year	MF SK						NBS					
	AE_{RO}	AE_{FO}	$RMSE_{RO}$	$RMSE_{FO}$	$MASE_{RO}$	$MASE_{FO}$	AE_{RO}	AE_{FO}	$RMSE_{RO}$	$RMSE_{FO}$	$MASE_{RO}$	$MASE_{FO}$
1995							1.0	0.2			1.0	*
1996							1.1	0.0			1.5	0.0
1997							0.8	0.2			0.4	*
1998							3.2	1.0			0.8	0.3
1999							0.8	0.1			0.6	1.0
1995 – 1999							1.4	0.3	1.6	0.5	0.9	0.4
2000							0.3	0.3			0.2	0.4
2001							0.9	0.7			0.7	0.4
2002							1.5	0.0			1.7	0.0
2003							1.3	0.9			7.9	0.9
2004	1.8	0.7			1.6	1.8	1.5	0.4			1.3	1.0
2000 – 2004	1.8	0.7	1.8	0.7	1.6	1.8	1.1	0.5	1.2	0.5	2.4	0.5
2005	3.1	2.8			1.5	1.0	3.0	2.7			1.4	0.9
2006	3.4	1.9			1.4	1.7	3.6	2.1			1.5	1.9
2007	1.2	0.5			0.2	0.2	1.8	0.2			0.4	0.1
2008	10.1	10.4			0.9	0.8	10.2	10.5			0.9	0.8
2009	3.2	2.2			0.3	0.2	2.0	1.0			0.2	0.1
2005 – 2009	4.2	3.5	5.2	5.0	0.9	0.8	4.1	3.3	5.2	5.0	0.9	0.8
2010	0.2	0.0			0.1	0.0	0.2	0.0			0.1	0.0
2011	0.1	1.0			0.1	2.5	0.8	0.3			0.6	0.8
2012	0.7	1.3			7.0	0.7	0.2	0.8			1.8	0.4
2013	0.3	0.4			0.3	0.2	0.3	0.4			0.3	0.2
2014	1.0	0.6			0.9	1.0	1.0	0.6			0.9	1.0
2010 – 2014	0.4	0.7	0.6	0.8	1.7	0.9	0.5	0.4	0.6	0.5	0.7	0.5
1995 – 2014	2.3	2.0	3.6	3.4	1.3	0.9	1.8	1.1	2.8	2.5	1.2	0.6

* Because of data attributes, MASE cannot be calculated for those years.

Source: Own calculations.

Slovak results largely correspond with the previous, regarding the main patterns. MF SK offered worse raw performance than the central bank (NBS), with the ministry faring slightly better vs. the naïve benchmark; interestingly, both institutions were outperformed by the naïve forecast in 2005 – 2009 and in the total FO period. As with Czech results, error measures peaked in economically breaking periods, most notably during accelerated growth in 2005 – 2006, following the outbreak of the global recession in 2008 and the EURO adoption in 2009. Error differences, again rather minor in size, point to an accuracy decrease when switching from FO to RO data. Similarly to the Czech Rep., one institution (MF SK) exhibited a gradual decrease of error differences over time, while the other one (NBS) exhibited an intermittent pattern. Overall, we can conclude that the accuracy measures in both countries do show some differences, between the forecasting errors originating from the RO and the FO underlying data. Most of the results also demonstrate that errors were higher with the RO than with the FO data, implying validity of the "forecaster aim for the first preliminary estimate" thesis. A crucial question now needs to be answered: are these visual differences also statistically significant?

Statistical Tests of Research Hypotheses

In the first part, a battery of two related-sample tests was carried out, in order to determine the statistical significance of the RO-FO error differences (H_1, H_2). This is the primary research topic of the paper and the corresponding results are quite interesting (Table 5, full details are provided in Appendix 2).

Table 5

RO-FO Error Differences Test Results (p-values)

	MFCR		**CNB**		**MFSK**		**NBS**	
	MAE (H_1)	MASE (H_2)	MAE (H_1)	MASE (H_2)	MAE (H_1)	MASE (H_2)	MAE (H_1)	MASE (H_2)
Related-samples Sign test	0.648	0.607	0.21	0.791	0.549	0.508	0.004[a]	0.049[a]
Related-samples Wilcoxon Signed Rank test	0.242	0.629	0.032[a]	0.638	0.142	0.676	0.001[a]	0.035[a]

[a] $p \leq 0.05$

Source: Own research.

According to both tests, the null hypothesis (no differences between the RO and the FO based error) was not rejected for two institutions completely (MFCR, MFSK), i.e. for both error measures (MAE, MASE). In other words, regarding these two entities, switching from the RO to the FO data does not significantly alter the scaled forecast accuracy, as well as the performance versus the naïve in-sample benchmark. On the other hand, we have the National Bank of Slovakia (NBS). Here, both tests rejected a null hypothesis with both error measures, suggesting that in this case the RO/FO interchange lead to a significant variation (decrease) of forecasting errors´ value. Finally, the Czech National Bank (CNB) forecasts exhibit an interesting, heterogeneous pattern. While with the MASE error, the null hypothesis (H_2) was not rejected, the MAE error produced conflicting results (H_1): the Sign test upheld the null hypothesis, yet Wilcoxon t. rejected it. This implies that the effect´s significance is not unequivocal in this case, but because we used Wilcoxon t. as our primary method, the interpretation should point rather to a significantly differentiating effect of the Y_t data vintage.

The second step of the analysis deals with differences of accuracy between individual institutions, under different Y_t data set-up (H_3 – H_6). Although not a direct examination, this analysis indicates, whether the data change affected the relative accuracy inside the sample. Kruskal-Wallis (K-W) and Mann-Whitney (M-W) U tests were utilised (Table 6, full details are provided in Appendix 3).

Table 6

Error Differences between Surveyed Institutions in Different Data Set-up (RO-FO; p-values)

Method	Whole sample					
	MAE			MASE		
Kruskal-Wallis test_RO data (H_3, H_4)	0.947			0.931		
Kruskal-Wallis test_FO data (H_5, H_6)	0.028[a]			0.178		
	MFCR x CNB	MFCR x MFSK	MFCR x NBS	CNB x MFSK	CNB x NBS	MFSK x NBS
Mann-Whitney U test_RO data_MAE (H_3)	0.951	0.435	0.646	0.906	0.509	0.757
Mann-Whitney U test_RO data_MASE (H_4)	0.626	0.968	0.871	0.655	0.401	0.635
Mann-Whitney U test_FO data_MAE (H_5)	0.279	0.795	0.010[a]	0.962	0.097	0.075
Mann-Whitney U test_FO data_MASE (H_6)	0.691	0.418	0.110	0.374	0.139	0.321

Source: Own research.

How do the results relate to the previous part? Firstly, regarding the whole sample differences (K-W test), both error measures offered different outcomes: while with the MAE, the RO/FO data switch did affect the statistical difference between institutions´ accuracy (RO null not rejected, FO null rejected), MASE results were identical in both setups (null not rejected). This can be seen as an implication that "raw" accuracy might be more sensitive to Y_t data than the comparison with the naïve benchmark. Furthermore, it is vital to note that in three cases (RO_MAE, RO_MASE, FO_MASE) the null hypotheses $H_{3.}$ H_4 and H_6 were not rejected, indicating no significant difference between sample members in terms of forecasting accuracy.

Regarding the bilateral differences (M-W test), final p-values paint a more homogeneous picture. Only in a single case out of 24 tested the interchange of the RO and the FO data resulted in diverse results (MFCR x NBS_{MAE}).[7] In other words, in the vast majority of the comparisons, change in the underlying Y_t data did not alter the relative accuracy among the surveyed institutions. In those cases, the H_3 – H_6 (null) hypotheses were not rejected, confirming that there is no statistically significant difference between forecasting the accuracy of affected institutions, in terms of the MAE and MASE errors. As with the K-W test, this is a vital secondary finding itself.

[7] Additional two MAE combinations (CNB x NBS, MFSK x NBS) could be considered on p = 0.1.

Discussion

Going through all of the analytical steps we have undertaken, study results have transpired through an important evolution. The first two steps, summary of different data vintages and visual comparison of the related error measures, have detected substantial variations in the surveyed data-lines and notable differences between forecasting errors originating with the RO and the FO data. Unsurprisingly, a vast majority of differentials points to an increase of forecasting error, when switching from the first to the recent out-turn Y_t vintage. Crucial is not the visual observation, however, but the following significance testing. And that confirmed that in a substantial majority of cases, the underlying RO/FO data switch neither caused significant differences in the MAE and the MASE error of individual institutions (Wilcoxon Signed Rank t.), nor changed the relative difference between the individual institutions´ accuracy (K-W t., M-W t.). This creates an interesting schism that can be simply explained, though: while the RO/FO data change does alter the accuracy of public growth forecasts in comprehensive directions, the magnitude of the changes is (in most cases) not high enough to cause significant differences. Procedurally, this evolution also underlines the importance of statistical testing as a critical element in forecasting analysis, in conjunction with plentiful evidence presented in literature (e.g. Hibon et al., 2012; Fildes and Steckler, 2000; Armstrong, 2001 and others).

From a theoretical perspective, study outcomes strongly support evidence presented in most concurrent studies outlined in the introduction part (Robertson and Tallman, 1998; Croushore and Stark, 2001; Aruoba, 2008 and others), listing only minor (insignificant) changes between the RO and the FO forecasting accuracy. This represents a continuation of doubting of claims raised by older works (McNees and Ries, 1983; Kenen and Schwarz, 1986), which were generally assuming critical importance of the RO/FO data input. The results, however, still imply that this assumption might hold true, if the data revisions present between the FO and the RO rise over a critical limit. While we have gained insignificant outcome in a strong majority of the cases (Wilcoxon sign t., M-W t.), there are partial results in certain years that do support this thesis (NBS Wilcoxon t. score, underlined by its MAE/MASE values in 2005 – 2009 period). This suggests interpretation conditions for our meritory conclusions, which is tied to a range of revisions between the FO and the RO. Should this exogenous parameter rise notably above the level inherent to current data (Table 2), in a manner described by for example Aruoba (2008), reassessment of the current findings is viable.

Regarding the practitioner's point of view, the conclusions of our paper can potentially bring a certain level of simplification to forecast assessment. While the evaluation based on the recent out-turn data can be expected to put more pressure on the institution´s performance, as it was found to be related to higher overall error, it should not differ, in general, significantly from an evaluation based on the first out-turn. This is an important message for the assessment methodologies different institutions and researchers use, as it weakens the necessity of “stick to the FO or the RO data” philosophies, which are sometimes rigidly enforced. It also broadens the manoeuvre space, when desired vintage of the data is unavailable or inconsistent. Secondly, the findings remain rather sceptical of the “forecasters aiming at the first available GDP estimate” thesis, represented by Di Fonzo (2005) or McKenzie and Gamba (2008). The scope of our research did not include differentiation between multiple early growth estimations (first available, settled etc.) at first, and as mentioned repeatedly: while their raw error measures indicate some difference, core test batteries disprove it. This promotes further examination of the issue in the follow up studies.

From an economic perspective, results need to be interpreted in a broader context of macroeconomic development of surveyed countries. Aside from business cycles effects, both economies had to deal with substantial interference coming from their fiscal and monetary policies. Previously unaccounted actions, like the CNB 2013 koruna exchange rate intervention, likely affected predictability of most macroeconomic indicators including GDP.[8] So did the arrival of the 2008 – 2009 financial crises, which is clearly reflected by the individual error measures (MAE, RMSE and MASE). In theory, the recent disruptive events of this type could be partly compensated by development of the forecasting methods themselves, which have over the years typically progressed into umpteenth iteration of Dynamic stochastic general equilibrium (DSGE) causal model adjusted by expert judgment. Both factors need to be taken into account when overseeing study outcomes.

Conclusions

The objective of this paper was to provide empirical evidence on the effect different data inputs (first vs. recent out-turn data) have on GDP growth forecast accuracy. Our subsequent findings can be summarised in the three main bullet points below:

- Summary statistics analysis revealed that, indeed, different vintage of data are subject to variation, resulting from continuous revisions. As evinced by the following analysis of forecasting error measures, this variation translates into lower accuracy of growth forecasts, when evaluated by the most recent (revised)

[8] For further details about macroeconomic development of both countries, consult e.g. periodic OECD Economic Outlook (OECD, 2016).

instead of the first available Y_t data (first estimate). While with the scaled measures (MAE, RMSE) the extent of difference generally reached individual units of error on average, the differences in performance vs. in-sample naïve benchmark (MASE) were considerably more substantial.

• In the statistical part of our analysis, nevertheless, the applied battery of tests (Sign t., Wilcoxon signed ranked t.) has indicated that the exchange of the underlying RO and FO data does not influence the accuracy of the forecasts in a significant way. Except for the NBS, with all other institutions, the null hypothesis (H_1, H_2) was rejected, for both error measures MAE and MASE.

• Our ultimate research inquiry was whether the change in the underlying Y_t data does affect the relative accuracy of individual institutions, i.e. previously insignificant differences in accuracy are made significant and vice versa. Utilising the Kruskal-Wallis and Mann-Whitney U tests, we have again found that in a strong majority of the cases, the change in the underlying data did not produce a difference between the RO/FO comparisons on selected significance levels, rejecting the null (H_3, H_4, H_5, H_6). The only notable exception was the whole sample MAE result, where the FO outcome indicated accuracy-difference, while the RO outcome disproved it. This was not reproduced in the paired testing (M-W t.), however.

The results of our research bring vital information for fellow researches and also both countries´ forecasting institutions and policymakers. Regarding the directions for future research, the study points out two particularly important areas. The first is obviously geographical extension – as the results encompass the area of the Czech and Slovak Republics, their verification in the broader regional context (of Central and Eastern, prospectively Western Europe) is vital. The second point is related to functional diversification, i.e. to reproduction of presented analysis on different macroeconomic indicators and their forecasts. Particularly inflation, due to its role in fiscal and monetary policy, represents an important candidate, with other primary aggregates (e.g. unemployment, GDP components) being suitable as well. Both directions would open the possibility for further generalization of our findings and should be also seen as general validity verification, required with any empirical survey.

As of study limitations, the main inherent parameter is the length of the time-scales available. This has objective as well as subjective reasons, with no credible macro-forecasts before 1995 (because of a system change in 1989) and some institutions exhibiting an even shorter forecasting history (MF SK, CNB). The extent of the time-scales used, however, is still comparable to similar studies reviewed in the theoretical part (i.e. Diron, 2008; Raponi and Frane, 2014; Rusnák, 2016), suggesting that data compliance assumption holds. Still, the verification of the results over time, using a lengthened time frame, is viable, in addition to the research opportunities listed above.

References

ALLAN, G. (2013): Evaluating the Usefulness of Forecasts of Relative Growth. [Working Paper, No. 12 – 14/2012.] Strathclyde: Strathclyde Business School.

ANTAL, J. – HLAVÁČEK, M. – HORVATH, R. (2008): Do Central Bank Forecast Errors Contribute to the Missing of Inflation Targets? The Case of the Czech Republic. Czech Journal of Economics and Finance (Finance a úvěr), *58*, No. 9 – 10, pp. 434 – 453.

ANTONIČOVÁ, Z. – MUSIL, K. – RŮŽIČKA, L. – VLČEK, J. (2009): Evaluation of the CNB's Forecasts. Economic Research Bulletin, *7*, No. 1, pp. 8 – 10.

ARMSTRONG, J. S. – COLLOPY, F. (1992): Error Measures for Generalizing about Forecasting Methods: Empirical Comparisons. International Journal of Forecasting, *8*, No. 1, pp. 69 – 80.

ARMSTRONG, J. S. (2001): Evaluating Forecasting Methods. In: ARMSTRONG, J. S. (ed.): Principles of Forecasting. Norwell: Kluwer Academic Publishers, pp. 443 – 472.

ARNOŠTOVÁ, K. – HAVRLANT, D. – RŮŽIČKA, L. – TÓTH, P. (2011): Short-Term Forecasting of Czech Quarterly GDP Using Monthly Indicators. Finance a úvěr – Czech Journal of Economics and Finance, *61*, No. 6, pp. 566 – 583.

ARTIS, M. J. (1996): How Accurate are the IMFs Short-Term Forecasts? Another Examination of the World Economic Outlook. [IMF Working Paper, No 96/89.] Washington, DC: International Monetary Fund.

ARUOBA, S. B. (2008): Data Revisions are not Well Behaved. Journal of Money, Credit and Banking, *40*, No. 2 – 3, pp. 319 – 340.

CROUSHORE, D. – STARK, T. (2001): A Real-time Data set for Macroeconomists. Journal of econometrics, *105*, No. 1, pp. 111 – 130.

CROUSHORE, D. – STARK, T. (2003): A Real-time Data set for Macroeconomists: Does the Data Vintage Matter? Review of Economics and Statistics, *85*, No. 3, pp. 605 – 617.

CZECH NATIONAL BANK (2016): Archiv prognóz a prezentací ze seminářů pro analytiky. Available at: <https://www.cnb.cz/cs/menova_politika/prognoza/predchozi_prognozy/index.html>.

CZECH STATISTICAL OFFICE (2016): HDP, národní účty. Available at: <https://www.czso.cz/csu/czso/hdp_narodni_ucty>.

DANÍELSSON, Á. (2008): Accuracy in Forecasting Macroeconomic Variables in Iceland. [Working Papers, No. 39.] Reykjavik: The Central Bank of Iceland.

DI FONZO, T. (2005): The OECD Project on Revisions Analysis: First Elements for Discussion. In: Ponencia presentada en la OECD steseg Meeting [Proceeding.] Barcelona: Organisation for Economic Co-operation and Development, pp. 27 – 28.

DIRON, M. (2008): Short-term Forecasts of Euro Area Real GDP Growth: An Assessment of Real-time Performance Based on Vintage Data. Journal of Forecasting, *27*, No. 5, pp. 371 – 390.

FAIR, R. C. (1986): Evaluating the Predictive Accuracy of Models. In: GRILICHES Z. – INTRILIGATOR, M. D. (eds): Handbook of Econometrics. Amsterdam: Elsevier Science Publisher, pp. 1979 – 1995.

FELDKIRCHER, M. – HUBER, F. – SCHREINER, J. – WOERZ, J. – TIRPAK, M. – TÓTH, P. (2015): Small-scale Nowcasting Models of GDP for Selected CESEE Countries [Working Paper, No. WP 4/2015.] Bratislava: Research Department, National Bank of Slovakia.

FENG, H. (2005): Real-time or Current Vintage: Does the Type of Data Matter for Forecasting and Model Selection? [Working Paper, No. EWP0515.] Victoria: University of Victoria, Department of Economics.

FILDES, R. – STECKLER, H. (2000): The State of Macroeconomic Forecasting. [Working Paper, No. 99-04.] Washington, DC: George Washington University, Centre for Economic Research.

GIANNONE, D. – REICHLIN, L. – SMALL, D. (2008): Nowcasting: The Real-time Informational Content of Macroeconomic Data. Journal of Monetary Economics, *55*, No. 4, pp. 665 – 676.

HARVEY, D. – LEYBOURNE, S. – NEWBOLD, P. (1997): Testing the Equality of Prediction Mean Squared Errors. International Journal of Forecasting, *13*, No. 2, pp. 281 – 291.

HIBON, M. – CRONE, S. – KOURENTZES, N. (2012): Statistical Significance of Forecasting Methods. [32nd Annual International Symposium on Forecasting.] Available at: <http://kourentzes.com/forecasting/wp-content/uploads/2014/04/ISF2012_Tests_Kourentzes.pdf>.

HYNDMAND, R. J. – KOEHLER, A. B. (2006): Another Look at Measures of Forecast Accuracy. International Journal of Forecasting, *22*, No. 4, pp. 679 – 688.

CHRISTENSEN, J. H. – DIEBOLD, F. X. – RUDEBUSCH, G. – STRASSER, G. (2007): Multivariate Comparisons of Predictive Accuracy. [Working Paper, No. 8/20.] Philadelphia: University of Pennsylvania.

KEEREMAN, F. (1999): The Track Record of the Commission Forecasts. [Working Paper, No. 137.] Brussels: Directorate General Economic and Monetary Affairs (DG ECOFIN), European Commission.

KENEN, P. B. – SCHWARTZ, S. B. (1986): An Assessment of Macroeconomic Forecasts in the International Monetary Fund's World Economic Outlook. [Working Paper.] Princeton: Princeton University, Department of Economics, International Finance Section.

LJUNG, G. M. – Box, G. E. (1978): On a Measure of Lack of Fit in Time Series Models. Biometrika, *65*, No. 2, pp. 297 – 303.

MAKRIDAKIS, S. – HIBON, M. (1995): Evaluating Accuracy (or Error) Measures. [Working Paper.] Fontainebleau: Institut Européen d'Administration des Affaires (INSEAD).

MANN, H. B. – WHITNEY, D. R. (1947): On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics, *18*, No. 1, pp. 50 – 60.

McKENZIE, R. – GAMBA, M. (2008): Interpreting the Results of Revision Analyses: Recommended Summary Statistics. [Working Paper.] Brussels: OECD/Eurostat Task Force on Performing Revisions Analysis for Sub-Annual Economic Statistics.

McNEES, S. K. – RIES, J. (1983): The Track Record of Macroeconomic Forecasts. New England Economic Review, *18*, No. 5, pp. 25 – 42.

MIAO, W. – GEL, Y. L. – GASWIRTH, J. L. (2006): A New Test of Symmetry about an Unknown Median. In: HSIUNG A. C. – YING Z. – ZHANG C. H. (eds): Random Walk, Sequential Analysis and Related Topics – A Festschrift in Honor of Yuan-Snih Chow. Singapore: World Scientific Publisher, pp. 199 – 214.

MINISTRY OF FINANCE OF THE CZECH REPUBLIC (2016): Makroekonomická predikce. Available at: <http://www.mfcr.cz/cs/verejny-sektor/makroekonomika/makroekonomicka-predikce>.

MINISTRY OF FINANCE OF THE SLOVAK REPUBLIC (2016): Makroekonomické prognózy. Available at: <http://www.finance.gov.sk/Default.aspx?CatID=112>.

NATIONAL BANK OF SLOVAKIA (2016): Strednodobá predikcia. Available at: <http://www.nbs.sk/sk/publikacie/strednodoba-predikcia>.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (2016): OECD Economic Outlook Archive. Available at: <https://stats.oecd.org/index.aspx?queryid=51396>.

ÖLLER, L. – BAROT, B. (2000): The Accuracy of European Growth and Inflation Forecasts. International Journal of Forecasting, *16*, No. 3, pp. 293 – 315.

RAPONI, V. – FRALE, C. (2014): Revisions in Official Data and Forecasting. Statistical Methods & Applications, *23*, No. 3, pp. 451 – 472.

ROBERTSON, J. C. – TALLMAN, E. W. (1998): Data Vintages and Measuring Forecast Model Performance. Economic Review-Federal Reserve Bank of Atlanta, *83*, No. 4, pp. 4 – 20.

RUSNÁK, M. (2016): Nowcasting Czech GDP in Real Time. Economic Modelling, *54*, Issue C, pp. 26 – 39.

STATISTICAL OFFICE OF THE SLOVAK REPUBLIC (SOSR) (2016): Národné účty. Available at: <https://slovak.statistics.sk/wps/portal/ext/themes/macroeconomic/accounts/>.

TAYLOR, N. (2014): Economic Forecast Quality: Information Timeliness and Data Vintage Effects. Empirical Economics, *46*, No. 1, pp. 145 – 174.

WILCOXON, F. (1945): Individual Comparisons by Ranking Methods. Biometrics, *1*, No. 6, pp. 80 – 83.

Appendices

Appendix 1

Independence and Symmetry Tests Results (p-values)

	Ljung-Box independence test[9]	Miao, Gel and Garswith symmetry test
MFCZ_MAE_ recent out-turn	0.525	0.919
MFCZ_MAE_ first out-turn	0.677	0.654
MFCZ_MASE_recent out-turn	0.679	0.588
MFCZ_MASE_ first out-turn	0.615	0.131
CNB_MAE_ recent out-turn	0.323	0.927
CNB_MAE_ first out-turn	0.351	0.740
CNB_MASE_recent out-turn	0.747	0.789
CNB_MASE_ first out-turn	0.505	0.609
MFSK_MAE_ recent out-turn	0.976	0.152
MFSK_MAE_ first out-turn	0.863	0.275
MFSK_MASE_recent out-turn	0.773	0.747
MFSK_MASE_ first out-turn	0.796	0.953
NBS_MAE_ recent out-turn	0.897	0.508
NBS_MAE_ first out-turn	0.943	0.594
NBS_MASE_recent out-turn	0.925	0.815
NBS_MASE_ first out-turn	0.891	0.967

Source: Own calculations.

Appendix 2

Sign t. and Wilcoxon Signed t. test Statistics (critical values for p = 0.05 in brackets)

	MFCR		CNB		MFSK		NBS	
	MAE (H_1)	MASE (H_2)	MAE (H_1)	MASE (H_2)	MAE (H_1)	MASE (H_2)	MAE (H_1)	MASE (H_2)
Related-samples Sign test *(k test statistics, binomial distribution)*	8 (6)	8 (6)	5 (5)	7 (5)	4 (2)	3 (2)	3 (6)	5 (6)
Related-samples Wilcoxon Signed Rank test *(V test statistics, binomial distribution)*	73 (53)	92 (53)	32 (35)	66 (35)	16 (11)	28 (11)	22 (53)	49 (53)

Source: Own calculations.

[9] Number of lags was set equal to ln(*n*).

Appendix 3

Kruskal-Wallis t. and Mann-Whitney U t. Test Statistics (critical values for p = 0.05 in brackets)

Method	**Whole sample**					
	MAE			MASE		
Kruskal-Wallis test_RO data (H_3, H_4) *(chi2 test statistics, chi2 distribution)*	0.367 (7.815)			0.444 (7.815)		
Kruskal-Wallis test_FO data (H_5, H_6) *(chi2 test statistics, chi2 distribution)*	9.099 (7.815)			4.917 (7.815)		
	MFCR x CNB	MFCR x MFSK	MFCR x NBS	CNB x MFSK	CNB x NBS	MFSK x NBS
Mann-Whitney U test_RO data_MAE (H_3) *(W test statistics, W distribution)*	168 (106)	91 (63)	183 (128)	91 (52)	148 (106)	102 (63)
Mann-Whitney U test_RO data_MASE (H_4) *(W test statistics, W distribution)*	154 (106)	109 (63)	194 (128)	84 (52)	142 (106)	98 (63)
Mann-Whitney U test_FO data_MAE (H_5) *(W test statistics, W distribution)*	134 (106)	104 (63)	106 (128)	92 (52)	115 (106)	67 (63)
Mann-Whitney U test_FO data_MASE (H_6) *(W test statistics, W distribution)*	157 (106)	90 (63)	141 (128)	74 (52)	121 (106)	86 (63)

Source: Own calculations.