Hinterlang, Natascha

Conference Paper

Predicting Monetary Policy Using Artificial Neural Networks

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2019: 30 Jahre Mauerfall - Demokratie und Marktwirtschaft - Session: Econometrics - Forecasting I, No. B07-V3

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Hinterlang, Natascha (2019) : Predicting Monetary Policy Using Artificial Neural Networks, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2019: 30 Jahre Mauerfall - Demokratie und Marktwirtschaft - Session: Econometrics - Forecasting I, No. B07-V3, ZBW - Leibniz-Informationszentrum Wirtschaft, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/203503

Predicting Monetary Policy Using Artificial Neural Networks*

Natascha Hinterlang†

Deutsche Bundesbank‡ and Goethe University Frankfurt

February 14, 2019

Abstract

This paper analyses the forecasting performance of monetary policy reaction functions using U.S. Federal Reserve's Greenbook real-time data. The results indicate that artificial neural networks are able to predict the nominal interest rate better than linear and nonlinear Taylor rule models as well as univariate processes. While in-sample measures usually imply a forward-looking behaviour of the central bank, using nowcasts of the explanatory variables seems to be better suited for forecasting purposes. Overall, evidence suggests that U.S. monetary policy behaviour between 1987-2012 is nonlinear.

*This paper is part of the authors' Ph.D. thesis supervised by Prof. Dr. Uwe Hassler. The author thanks him, Prof. Dr. Matei Demetrescu, Dr. Christian Speck and Marc Pohle for many helpful comments.

†Deutsche Bundesbank, DG Economics, Public Finance Division, Wilhelm-Epstein-Strasse 14, 60431 Frankfurt am Main, Germany, Natascha.Hinterlang@hof.uni-frankfurt.de

‡The views expressed in this paper are those of the author; they do not necessarily reflect the views of the Deutsche Bundesbank or its staff.

1 Introduction

The federal funds rate as the main monetary policy instrument of the Fed is the most important interest rate for the U.S. economy. Changes in the policy rate can have real effects via different transmission channels, e.g. the interest rate, the asset price and wealth, and the exchange rate channel. Hence, the federal funds rate is of particular interest for various market participants. In order to base today's decisions affecting future outcomes (investments, bank lending etc.) on it, reliable forecasts of the federal funds rate are necessary. Moreover, the Fed itself is interested in how people form expectations about monteary policy since forward-guidance provides an additional tool for central banks to reach its inflation and/ or output targets. To facilitate markets' expectation formation, central bankers often refer to the extraction of their reaction function from the data[1].
Hence, many studies were undertaken to find the most appropriate form. The idea of a rules-based monetary policy dates back to Taylor (1993). He proposed a reaction function where the nominal interest rate depends linearly on the gaps between actual and targeted values of inflation and output. This simple "Taylor rule" was shown to match federal reserve actual interest setting behaviour between 1987 and 1992 very well. Moreover, Svensson (1997) provided a theoretical model which supports a linear reaction function as the solution to a central banker's optimization problem of minimizing the deviations of inflation and output from their desired values. However, the result depends on the assumptions of a quadratic loss function reflecting symmetric preferences and a linear Phillips curve. Clarida et al. (1998, 2000) added lagged values of the federal funds rate explained by an interest rate smoothing motive and manifested the forward-looking version of the Taylor rule, in which the central bank focuses on expected inflation and output gap instead of past or current values. Assuming asymmetric preferences (see e.g. Nobay and Peel (2003); Cukierman and Muscatelli (2008); Ruge-Murcia (2004)), a convex aggregate supply curve (Schaling (2004); Dolado et al. (2004)) or both (Dolado et al. (2005)) yields a nonlinear policy rule. The majority of studies on nonlinear Taylor rules employs smooth transition models (amongst others) to capture asymmetric preferences of the central bank. While Kim et al. (2005), Qin and Enders (2008) and Castro (2011) do not find evidence of nonlinearities over the periods 1979-2000, 1987-2005 and 1982-2007, respectively, Petersen (2007) suggests asymmetric behaviour of the Fed depending on the level of inflation during the period 1985-2005.
The paper at hand is closely related to these studies in the sense that its model compari-

[1] See e.g. Yellen (2017) and Draghi (2018)

son includes smooth transition functions with different activation functions and threshold specifications. Moreover, it extends the nonlinear model list by so-called artifical neural networks (ANNs). The ANN has the property of being an universal approximator, i.e. it can fit in-sample data to any degree. Even though it does not provide a structural interpretation of the estimated parameters, ANNs can serve as a useful forecasting tool for time series[2]. Hence, the paper focuses on comparing (pseudo-) out-of-sample performances across models, which also accounts for the fact that nonlinear models are prone to overfitting problems. To my knowledge, there exists only one paper by Malliaris and Malliaris (2009) considering ANNs in the context of monetary policy reaction functions. They find that the ANN outperforms a linear Taylor rule and a random walk only when the data is split based on the current value of the federal funds rate, but not when it is time-based. However, their forecasted periods are randomly drawn from subsamples, which lacks a realistic forecasting simulation in a time-series context. This paper provides that analysis by performing expanding window regressions and explicitly considering only information that was available at the time the forecast was made. Orphanides (2001) already emphasized the importance of using real-time data in a Taylor rule framework since inflation and output gap *ex post* measures might be different due to revision processes yielding misleading reaction functions.

This paper contributes to the literature in three ways. First, by introducing artifical neural networks with a clear specification scheme, it adds a powerful tool for forecasting monetary policy. Second, thanks to better real-time data availabilty, forecasts are solely based on data that were in the Fed's information set at that time. Thus, it provides a more realistic forecasting situation than comparable studies. Third, it offers additional evidence for a nonlinear reaction function for the era since Greenspan.

The paper is organized as follows. Section 2 introduces the different model specifications considered in the empirical analysis and describes the data. In Section 3 the models' pseudo-out-of-sample forecasting performance is compared including Diebold-Mariano forecast accuracy tests. Section 4 discusses the results and Section 5 concludes.

2 The Models and Data

This section presents all model specifications used for forecasting the federal funds rate. The models differ in two dimensions - the functional form (i.e. linear, smooth transition or

[2]See e.g. Teräsvirta et al. (2005) and Gonzalez (2000) for succesful applications of ANNs in a macroeconomic time-series forecasting context.

ANN) and the input dimension (i.e. within-quarter, backward- or forward-looking). The latter refers to the timing of the exogenous variables inflation and output gap. While the within-quarter (WQ) specification uses nowcasts of inflation and the output gap given the information set at time t, i.e. $\pi_{t|t}$ and $y_{t|t}$, respectively, the backward-looking (BW) version includes lagged values, i.e. $\pi_{t-1|t}$ and $y_{t-1|t}$. The forward-looking (FW) specification uses the one-quarter-ahead forecasts $\pi_{t+1|t}$ and $y_{t+1|t}$. Besides the linear and nonlinear models, the comparison also includes two univariate processes. The employed data is described at the end of this section.

2.1 Linear Models

The linear models are all modifications of the original version of the Taylor (1993) rule including policy inertia in the lines of Clarida et al. (1998) with two lagged interest rate terms. They only differ in the timing of the explanatory variables inflation and the output gap, i.e. π and y, respectively.

$$\textbf{Linear-WQ:} \quad i_t = (1-\rho)(\alpha + \theta\pi_{t|t} + \beta y_{t|t}) + \rho_1 i_{t-1} + \rho_2 i_{t-2} + \varepsilon_t \tag{1}$$

$$\textbf{Linear-BW:} \quad i_t = (1-\rho)(\alpha + \theta\pi_{t-1|t} + \beta y_{t-1|t}) + \rho_1 i_{t-1} + \rho_2 i_{t-2} + \varepsilon_t \tag{2}$$

$$\textbf{Linear-FW:} \quad i_t = (1-\rho)(\alpha + \theta\pi_{t+1|t} + \beta y_{t+1|t}) + \rho_1 i_{t-1} + \rho_2 i_{t-2} + \varepsilon_t \tag{3}$$

with $\rho = \rho_1 + \rho_2$. Model (1) uses the nowcast of the explanatory variables, while model (2) includes previous period's values of inflation and the output gap. In model (3), the interest rate depends on one-quarter ahead expected values of both.[3]

2.2 Nonlinear Models

Smooth Transition Models

Smooth transition (STR) models are one of the most popular nonlinear models and mostly used in the context of nonlinear Taylor rules since they allow for regime-switching (asymmetric) central bank behaviour. The general structure is defined as follows:

$$i_t = \alpha_0 + \alpha_1\pi_t + \alpha_2 y_t + \alpha_3 i_{t-1} + \alpha_4 i_{t-2} + G(\gamma, c, s_t)\cdot(\beta_0 + \beta_1\pi_t + \beta_2 y_t + \beta_3 i_{t-1} + \beta_4 i_{t-2}) + \varepsilon_t.$$

[3] All models are estimated by nonlinear least squares (NLLS) with HAC standard errors (Bartlett Kernel, Newey-West fixed bandwidth) in EViews 10.

It consists of a linear part $\alpha_0+\alpha_1\pi_t+\alpha_2 y_t+\alpha_3 i_{t-1}+\alpha_4 i_{t-2}$ and a nonlinear part $G(\gamma, c, s_t)\cdot(\beta_0+\beta_1\pi_t+\beta_2 y_t\beta_3 i_{t-1}+\beta_4 i_{t-2})$, where $G(\gamma, c, s_t)$ is a continuous and bounded (between 0 and 1) transition function, with slope γ, threshold parameter(s) c and a transition variable s_t. The slope γ, also known as the smoothness parameter, indicates the speed of the transition from 0 to 1. Transition functions considered in this paper are the logistic (LSTR), the logistic-quadratic (L2STR) and the exponential (ESTR):

$$\textbf{LSTR-WQ:} \quad i_t = \alpha_0+\alpha_1\pi_{t|t}+\alpha_2 y_{t|t}+\alpha_3 i_{t-1}+\alpha_4 i_{t-2}+ \{1+exp[-\gamma(s_t-c)]\}^{-1}(\beta_0+\beta_1\pi_{t|t}+\beta_2 y_{t|t}+\beta_3 i_{t-1}+\beta_4 i_{t-2})+\varepsilon_t \tag{4}$$

$$\textbf{ESTR-WQ:} \quad i_t = \alpha_0+\alpha_1\pi_{t|t}+\alpha_2 y_{t|t}+\alpha_3 i_{t-1}+\alpha_4 i_{t-2}+ \{1-exp[-\gamma(s_t-c)^2]\}(\beta_0+\beta_1\pi_{t|t}+\beta_2 y_{t|t}+\beta_3 i_{t-1}+\beta_4 i_{t-2})+\varepsilon_t \tag{5}$$

$$\textbf{L2STR-WQ:} \quad i_t = \alpha_0+\alpha_1\pi_{t|t}+\alpha_2 y_{t|t}+\alpha_3 i_{t-1}+\alpha_4 i_{t-2}+ \{1+exp[-\gamma(s_t-c_1)(s_t-c_2)]\}^{-1}(\beta_0+\beta_1\pi_{t|t}+\beta_2 y_{t|t}+\beta_3 i_{t-1}+\beta_4 i_{t-2})+\varepsilon_t \tag{6}$$

with $\gamma \geq 0$. The logistic transition function of model (4) is monotonically increasing in the threshold variable s_t. Hence, the central banks reacts differently for high and low values of s_t representing asymmetric preferences. Model (5) relies on an exponential transition function, that is increasing in absolute deviations of s_t from the c, indicating symmetric behaviour around the point $s_t = c$. For $\gamma \to 0$, both models become linear. While for $\gamma \to \infty$, the LSTR model approaches the discrete 2-regime threshold model, the ESTR model becomes linear since $G(\cdot) \to 1$. The L2STR model (6) nests a 3-regime discrete threshold model since $G(\cdot) \to 1$ for $s_t < c_1$ and $s_t > c_2$ and $G(\cdot) \to 0$ for s_t in-between. For $\gamma \to \infty$, it becomes linear. The L2STR model allows the central bank to target a band instead of a single point of the threshold variable.

Note that models (4)-(6) are stated in the within-quarter version. The backward- and forward-looking versions, where $\{\pi_{t|t}, y_{t|t}\}$ is replaced by $\{\pi_{t-1|t}, y_{t-1|t}\}$ and $\{\pi_{t+1|t}, y_{t+1|t}\}$, respectively, are considered as well in the empirical analysis. Moreover, for all specifications the threshold variable is allowed to be either inflation ($\pi_{t|t}/\pi_{t-1|t}/\pi_{t+1|t}$) or the output gap ($y_{t|t}/y_{t-1|t}/y_{t+1|t}$) since both variables are targeted by the Fed[4].

Artificial Neural Networks

The idea of ANNs as an application to artificial intelligence already dates back to the 1940s. However, it has become more popular in the late 90s due to the massively increased

[4]Results are only reported for the threshold variable with the better forecasting performance. It is not necessarily the one that minimizes the residual sum of squares in-sample.

processing power of computers (see e.g. Haykin (1999)). The ANN considered in this paper is a so-called "single-hidden-layer" recurrent neural network with the following form:

$$i_t = \alpha_0 + \sum_{j=1}^{q} \gamma_j G(\boldsymbol{\beta_j'} \mathbf{z_t} + \alpha_j) + \varepsilon_t, \tag{7}$$

where $\mathbf{z_t}$ is the vector of inputs, i.e. autoregressive and exogenous explanatory variables. The parameters to be estimated are $\boldsymbol{\beta_j}$ and γ_j, $j = 1, \ldots, q$ (also called "weights") and α_i, $i = 0, \ldots, q$ (also known as "biases"). Furthermore, $G(\cdot)$ is a bounded and monotonically increasing transfer function similar to the STR models. More specifically, the one used in this analysis is the hyperbolic tangent sigmoid function $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, that maps on the interval $[-1, 1]$.[5] The overall structure is described graphically by means of Figure 1.

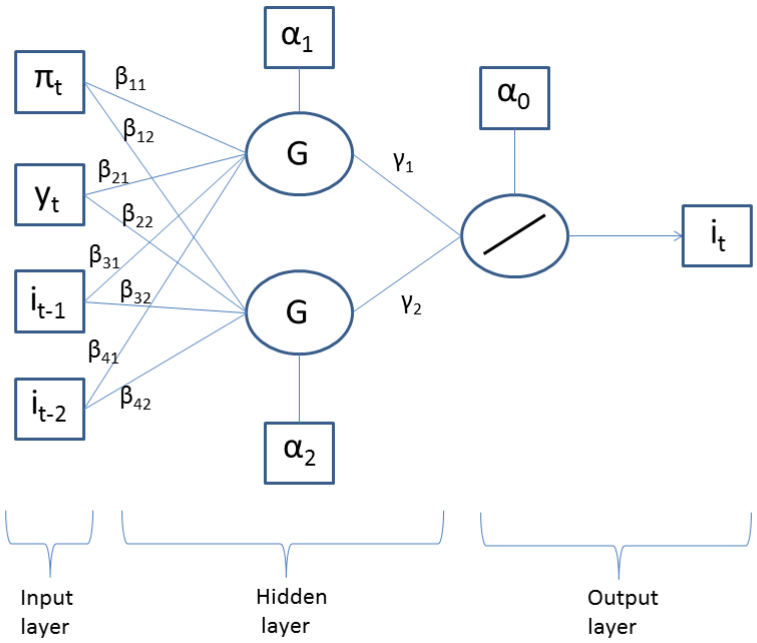

Figure 1: Single-hidden-layer neural network with $q = 2$ hidden units.

The "single-hidden-layer" neural network consists of three layers in total. The input layer contains the explanatory variables, while the dependent variable combined with a simple linear transfer function builds the output layer. The "hidden-layer" comprises q (here $q = 2$ for simplicity) so-called "hidden units" or "neurons" representing a transfer function G, where each is connected with all inputs via the weights β_{ij} (i.e. the connection from

[5]The hyperbolic tangent as a symmetric sigmoid function is often used because of faster convergence rates in comparison to the standard logistic function (see e.g. LeCun et al. (2012)).

input i to neuron j) . The output value of each neuron is the hyperbolic tangent function G evaluated at the value of the sum of the weighted inputs plus a bias term α_j. All neuron output values are then weighted again by γ_j and another bias term α_0 is added. Stuck together in the linear output transfer function yields one final value for the interest rate i_t. Consequently, the ANN-WQ (equivalently, the ANN-BW and ANN-FW) reaction function looks as follows:

$$\textbf{ANN-WQ:} \qquad i_t = \alpha_0 + \sum_{j=1}^{q} \gamma_j \cdot G(\alpha_j + \beta_{1j}\pi_{t|t} + \beta_{2j}y_{t|t} + \beta_{3j}i_{t-1} + \beta_{4j}i_{t-2}) + \varepsilon_t, \quad (8)$$

where $G(\cdot)$ denotes the hyperbolic tangent sigmoid transfer function. Since it is assumed that the federal funds rate depends on lags of itself as well as on inflation and the output gap, the network type considered here is the "nonlinear autoregressive network with exogenous inputs" (NARX).

One crucial task in estimating ANNs is the determination of the number of hidden units q. The strategy how it is chosen in this paper follows these steps: First, the estimation sample is split into a training and a validation set. In order to preserve the time series structure, the former consists of the first 80% and the latter of the last 20% of the observations. Second, the ANN is trained, i.e. estimated, looping over $q = 1, \ldots, 10$ hidden units with 30 different randomly chosen[6] initial weights and biases each[7]. Estimation is done in Matlab 2017Rb by the Levenberg-Marquardt algorithm (LMA), a non-linear least squares solver that combines the Gauss-Newton algorithm with the gradient descent method, together with an "early stopping" procedure. The latter ensures that training stops if the network performance, i.e. the mean squared error, fails to improve or remains the same for 6 consecutive epochs. Thus, the validation set only prevents overfitting and is not used for estimation. Third, the number of hidden units that minimizes the resulting mean squared error in the validation set averaged over the 30 trials is chosen to be the optimal one q_{opt}. Fourth, holding q_{opt} fixed, the trial with the lowest validation mean squared error serves as the optimal initial weights and biases.

The use of ANN models is motivated by the universal approximator property first shown by Hornik et al. (1989). It states that any unknown function H (under mild regularity assumptions) can be approximately arbitrarily close by a linear combination of activation functions G, i.e. $|H(z_t) - \sum_{j=1}^{q} \gamma_j G(\boldsymbol{\beta_j'} \boldsymbol{z_t})| < \delta$ with finite q and $\delta \in \mathbb{R}_{>0}$. Thus, the main advantage of ANNs is that one does not need to specify a specific functional form since

[6]The used Matlab default is the Nguyen-Widrow method for weights initialization. The seed was set to 100.
[7]See Aras and Kocakoç (2016) for a similar model selection strategy.

its specification is data-driven. However, it involves the drawbacks of being only locally identified and that parameters lack economic interpretation. Hence, its purpose is mainly forecasting, which suffices for the task of predicting monetary policy while neglecting deeper structural interpretations of the Fed's behaviour.

2.3 Univariate Models

Besides the multivariate models, two univariate models are included in the model comparison as well. Specifically, the AR(2) and ARIMA(1,1,0) are used for forecasting the federal funds rate[8].

$$\textbf{AR(2):} \quad i_t = \alpha_0 + \alpha_1 i_{t-1} + \alpha_2 i_{t-2} + \varepsilon_t \tag{9}$$

$$\textbf{ARIMA(1,1,0):} \quad \Delta i_t = \alpha_0 + \alpha_1 \Delta i_{t-1} + \varepsilon_t \tag{10}$$

For a list of all individual 17 model specifications considered in the forecasting comparison plus a forecast combination model see Table 8 in the Appendix.

2.4 The Data

The Federal Reserve Greenbook, which is produced in preparation of each meeting of the Federal Open Market Commitee, serves as the data source for inflation in this study. It provides real-time back-, now- and forecasts reaching from four quarters back up to nine quarters ahead. Moreover, there is a real-time data set on output gaps available, that was not included in the Greenbook, but was constructed and employed by the Board of Governors staff[9]. It covers projections of the output gap from eight quarters back up to nine quarters ahead. The use of these real-time data sets ensures that only information is used in the forecasts that was actually available by the Fed at the time they set the interest rate. It circumvents the potential problem of estimating misleading reaction functions due to the use of revised data as pointed out by Orphanides (2001). The sample covers the periods 1987:Q3-2012:Q4, where the starting period corresponds to the appointment of Alan Greenspan as the Fed's chairman, and the ending period is due to fact that the data is published only after a lag of five years. It covers a time span where inflation targeting was implicitly practiced (see Goodfriend (2004)) and explicitly defined (see Bernanke and Boivin (2003)).

[8] Experiments with different lag lengths did not improve the results.

[9] https://www.philadelphiafed.org/research-and-data/real-time-center/greenbook-data/gap-and-financial-data-set

The federal funds rate is obtained from FRED Economic Data and transformed to a quarterly average. The inflation rate is measured by the growth rate of the Core Consumer Price Index (CCPI), because the real-time data set on the Core Personal Comsumption Expenditure (CPCE) Price Index starts only in 2000:Q1.[10] The output gap is defined as the difference between actual and potential output expressed as a percentage of potential output. Concerning timing, data from the middle month (or, the first month if not available) of the respective quarter is applied.

3 Forecasting Results

In this section, the forecasting performance of the above described models is analysed. Specifically, expanding window regressions are employed, i.e. after estimating the first window covering the first T observations, data from $T+1$ is added for the estimation of the second window and so forth. After each window regression, one-, two-, three- and four-step-ahead forecasts are calculated. Thereby the recursive approach is used. For example, consider the case $k=3$: i_{t+3} depends on i_{t+2} and i_{t+1} due to the autoregessive structure inherent in all models. For the forecast $i_{t+3|t}$ however, i_{t+2} and i_{t+1} are replaced by the prior forecasts $i_{t+2|t}$ and $i_{t+1|t}$. With respect to the exogenous variables inflation and output gap, available real-time forecasts are substituted corresponding to the three input timing versions:

$$\begin{aligned}
\textbf{WQ:} \quad i_{t+k|t} &= f(i_{t+k-1|t}, i_{t+k-2|t}, \pi_{t+k|t}, y_{t+k|t}) \\
\textbf{BW:} \quad i_{t+k|t} &= f(i_{t+k-1|t}, i_{t+k-2|t}, \pi_{t+k-1|t}, y_{t+k-1|t}) \\
\textbf{FW:} \quad i_{t+k|t} &= f(i_{t+k-1|t}, i_{t+k-2|t}, \pi_{t+k+1|t}, y_{t+k+1|t}),
\end{aligned}$$

with $k = 1, \ldots, 4$.

The initial estimation window comprises the periods 1987:Q3-2000:Q2. Before a forecast is made, it is used to determine the number of hidden units in the artifical neural network in the way described in Section 2.2. The chosen numbers of hidden units are 4, 1 and 2 for the within-quarter, the backward-looking and the forward-looking version, respectively. Concerning the smooth transition models, inflation as well as the output gap are allowed to be the transition variable. Table 1 only reports results for the better out-of-sample

[10] The Fed's prefered inflation measure actually changed from CPI to PCE in 2000 and from PCE to CPCE in 2004. Orphanides and Wieland (2008) ,however, show that these switches of the inflation concept do not have a substantial effect on the reaction functions' estimates.

performing version. There are only four cases were the output gap is chosen to be the threshold variable. These are the within-quarter and the forward-looking logistic and exponential smooth transition models. For simplicity, all models are non-adaptive in the sense that their structures are not re-optimized after each estimation window; only the coefficients are re-estimated.

Since the actual values of the federal funds rate are available for all periods, this "pseudo-out-of-sample" analysis allows the computation and comparison of root mean squared forecast errors (RMSFE).

As explained in e.g. Teräsvirta et al. (2010), the multi-step-ahead forecasts from the STR and the ANN models cannot be obtained recursively. Therefore, the Monte Carlo method (see Teräsvirta et al. (2010, Ch. 14.2.2)) (with 1000 replications for each forecast) is employed for the multi-step-ahead forecasts from the STR models. However, the differences to their so-called "naive" method were neglectable. Hence, the ANN model forecasts rely on the "naive" method in order to keep computational burden low. Table 1 reports the forecasting performance results in terms of RMSFEs. It also includes the performance of the combined forecast, which is simply the mean of the individual forecasts[11].

First of all, looking at the RMSFEs, the within-quarter version of the artifical neural network outperforms all other models at all forecasting horizons. Especially, it dominates the univariate models and the combined forecasts as well. Interestingly, it's superiority is increasing in the length of the forecasting horizon. The backward-looking linear model performs better than the univariate specifications at horizons 3 and 4 and it has lower RMSFEs compared to the STR models at horizons 1-3. Within the class of STR models, the backward-looking version of the logistic-quadratic model performs best at horizons 1 and 2, while the within-quarter version of the exponential (with the output gap being the threshold variable) dominates for $k = 3, 4$. The within-quarter logistic and exponential STR model outperform the backward-looking linear model and the univariate ones at forecast horizon 4. However, they come off badly compared to the artifical neural network. The forward-looking version performs poorly over all functional forms and the backward-looking version seems to be particularly unsuited for the artificial neural network.

Table 2 summarizes the forecasting comparison results by reporting the average rank of each model over the four forecast horizons according to the RMSFE. The ANN-WQ model is ranked first for all forecasting horizons. The combined and univariate forecasts are listed on two to four with an average rank of five. The Linear-BW model follows in front of the

[11] Combining forecasts by taking the median forecast did not improve the result

Forecasting results (RMSFE)

Model		$k=1$	$k=2$	$k=3$	$k=4$
Linear	WQ	0.3994	0.7947	1.1877	1.5851
	BW	0.3807	0.7251	1.0807	1.4604
	FW	0.4506	0.9255	1.4057	1.8872
LSTR	WQ	0.3852	0.7427	1.0927	1.4527
	BW	0.3834	0.7514	1.1293	1.5496
	FW	0.3994	0.7619	1.1036	1.4488
ESTR	WQ	0.3836	0.7477	1.0945	1.4429
	BW	0.3853	0.7511	1.1425	1.5553
	FW	0.4072	0.7764	1.1352	1.5341
L2STR	WQ	0.3982	0.7572	1.1031	1.4471
	BW	0.3836	0.7385	1.1169	1.5202
	FW	0.4242	0.7948	1.1697	1.5579
ANN	WQ	0.3572	0.6579	0.9486	1.2540
	BW	0.5535	1.1508	1.7289	2.2494
	FW	0.4860	0.9176	1.2762	1.6042
AR(2)		0.3703	0.7120	1.0867	1.4936
ARIMA(1,1,0)		0.3692	0.7114	1.0916	1.5135
Mean		0.3667	0.6981	1.0316	1.3868

Table 1: Root mean squared forecasting errors (RMSFEs) for forecasting horizons $k = 1, \ldots, 4$ after expanding window regressions. The initial estimation period is 1987:Q3-2000:Q2. The structure of the WQ-/ BW-/ FW-ANN consists of 4, 1 and 2 hidden units, respectively. LSTR-WQ, ESTR-WQ, LSTR-FW and ESTR-FW use the output gap as the threshold variable; while all other STR models use inflation.

STR models. Surprisingly, there is no forward-looking model within the Top 10, although, literature shows broad consensus on monetary policy being forward-looking. However, the majority of these studies focuses on the comparison of in-sample fit measures. It seems to be the case, that the forward-looking version dominates in-sample[12], while being less suited for forecasting exercises. The fact that interest rate forecasts of the forward-looking model rely on forecasts of the exogenous explanatory variables that reach further

[12]This is also found in own in-sample comparisons, that are not reported here.

in the future compared to the within-quarter or backward-looking model could explain this phenomenom. While the backward-looking and the within-quarter models need $\pi_{t+k-1|t}$, $y_{t+k-1|t}$ and $\pi_{t+k|t}$, $y_{t+k|t}$, respectively in order to forecast $i_{t+k|t}$, the forward-looking model uses $\pi_{t+k+1|t}$ and $y_{t+k+1|t}, k = 1, \ldots, 4$. Using the within-quarter version or the backward-looking one might simply produce better interest rate forecasts due to smaller forecast errors on the input variables side.

Forecasting ranks

Model	Average rank
ANN-WQ	1
Mean	2
ARIMA(1,1,0)	5
AR(2)	5
LINEAR-BW	5
ESTAR-WQ	6.25
LSTAR-WQ	7
L2STAR-WQ	8.5
L2STAR-BW	8.5
LSTAR-BW	9.75
LSTAR-FW	9.75
ESTAR-BW	11.25
ESTAR-FW	12.5
LINEAR-WQ	14
L2STAR-FW	14.5
ANN-FW	16.25
LINEAR-FW	16.75
ANN-BW	18

Table 2: Average forecasting ranks over the forecasting horizons $h = 1, \ldots, 4$ according to the root mean squared forecasting error (RMSFE).

Figure 2 plots the forecasts from selected models (ANN-WQ, ARIMA(1,1,0), Linear-BW and ESTR-WQ) together with the actual federal funds rate. Subfigures a), b), c) and d) show the results for $k = 1, \ldots, 4$, respectively. For the lowest forecasting horizon ($k = 1$), all models perform quite well. The longer the horizon, the more distinguished are the forecasts. Interestingly, the ARIMA(1,1,0) and the Linear-BW model over-predict,

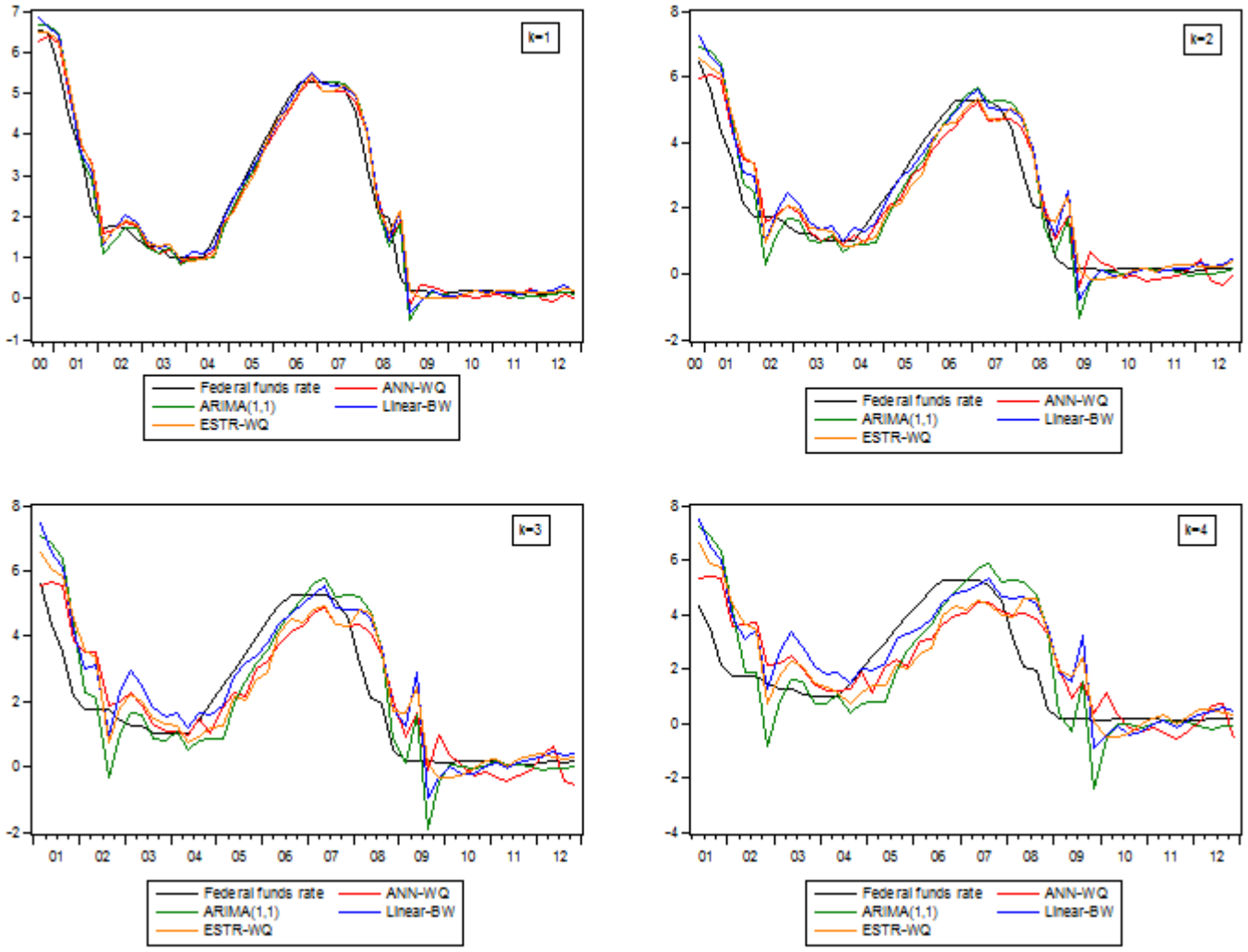

Figure 2: Forecasts of ANN-WQ, ARIMA(1,1,0), Linear-BW and ESTR-WQ and actual federal funds rate for forecasting horizons $k = 1, \ldots, 4$. .

while the nonlinear models under-predict the federal funds rate prior to the crisis drop beginning in 2007. Moreover, all models fail to incorporate the zero lower bound since they predict negative interest rates since 2009. For the ARIMA(1,1,0) model, the negative peak is especially large, while the ANN-WQ model only marginally falls below zero. The latter is also the only model that shortly over-predicts the federal funds rate between 2009-2010. Overall, the ANN-WQ is characterized by a smoother course with less over- and undershooting compared to the other models.

Modified Diebold-Mariano (MDM) Tests

Looking at the RMSFEs and the graphs, the ANN-WQ seems to be the best forecasting model for the federal funds rate. In the following, a test for equal predictive accuracy by Diebold and Mariano (1995) is employed in order to investigate the statistical significance of this result.

The approach consists of pairwise forecast comparisons testing the null hypothesis H_0 : $MSFE_m \geq MSFE_j$, i.e. model $m's$ forecasting performance is *not* superior to the one of model j. The test is based on the difference of squared errors $D_{mjt} = (i_t - f_t^{(m)})^2 - (i_t - f_t^{(j)})^2$. Here, the modified test statistic of Harvey et al. (1997) is employed:

$$MDM_k = \sqrt{\frac{T + 1 - 2k + \frac{1}{T}k\,(k-1)}{T}} \cdot \frac{\frac{1}{T}\sum_{t=1}^{T} D_{mjt}}{\sqrt{\frac{1}{T}\hat{\gamma}_D}}, \tag{11}$$

where T is the number of observations in the forecast series, k is the forecasting horizon and $\hat{\gamma}_D$ is an estimate of the long-run variance of D_{mjt}. p-values are taken from the Student's t-distribution with $(T-1)$ degrees of freedom to account for possible small sample size issues.

Actually, the Diebold-Mariano test was intended for comparing model-free forecasts as pointed out by Diebold (2015). Comparing econometric models via pseudo-out-of-sample forecasts complicates the test's asymptotics. West (2006) and Clark and McCracken (2001, 2013) show that the limiting distribution might be non-normal depending on the models' structures and estimation designs, e.g. nested or non-nested models and rolling/ expanding/ fixed estimation scheme. As the pairwise comparisons here include cases where models are nested (e.g. all STR models nest the univariate AR(2) model), their critique applies. However, Clark and McCracken (2011) find that standard normal critical values often approximate the precise distribution very well. Hence, I follow the lines of Diebold (2015) by sticking to the Gaussian limiting distribution and testing the validity of the sufficient assumption of covariance stationary loss differentials D_{mjt}.[13] Note, that a rejection of the null hypothesis of equal predicticy accuracy implies that it will also reject with an even smaller p-value if an asymptotic valid test is used (see West (2006, Table 3C, 1.b.)). Tables 9-12 in the Appendix report Augmented-Dickey-Fuller test results with the null hypothesis

[13] Following the exact route would require bootstrapping since the conditions for using critical values from McCracken (1999) are not met (see West (2006, Table 3C, 1.)).

of nonstationary loss differentials D_{mjt} for $k = 1, \ldots, 4$, respectively. The unit root null hypothesis is rejected at conventional significance levels for most of the pairwise model combinations. The cases where it cannot be rejected are of minor importance since the difference of the respective MSFEs is of such magnitude that it is clear which model is more appropriate. The general validity of the covariance stationary loss differentials assumption allows to continue with the modified Diebold-Mariano test. Since the ANN-WQ model has the lowest MSFEs, it is of particular interest, if it's forecast superiority is statistically significant. Therefore, Tables 3 and 4 report p-values of the modified Diebold-Mariano test with the ANN-WQ chosen as model m and j, respectively, for all forecasting horizons. The p-values of all other pairwise tests are listed in Tables 13-16 in the Appendix.
The MDM p-values support the first impression, that the ANN-WQ's forecasting superiority is increasing with the forecasting horizon. The number of pair-wise forecasting comparisons where the $MSFE_{ANN-WQ}$ is statistically lower at $\alpha = 10\%$ than the one of model j almost doubles (7 vs. 13) from $k = 1$ to 4. There are only four models, where the null hypothesis cannot be rejected at conventional significance levels, irrespective of the forecasting horizon. These are the Linear-BW, the AR(2), the ARIMA(1,1,0) models as well as the Mean forecast combination. Compared to these models' forecasts, the ANN-WQ's forecast is not statistically superior. However, if the null hypothesis is switched to $MSFE_m \geq MSFE_{ANN-WQ}$, Table 4 shows that it can never be rejected as well. Taking a look at the previously mentioned four models, we notice that the p-values are much larger compared to the respective p-values from Table 3. Hence, one would rather reject the null hypothesis that the ANN-WQ's forecast is *not* superior to the others than the other way around, speaking in favour of the ANN-WQ model.

In order to obtain an overall significance level, the $N = 18$ individual null hypothesis are combined to an overall null hypothesis $H_0 : H_{0,1} \cap H_{0,2} \cap \ldots \cap H_{0,18}$. The idea of this p-value combination dates back to Fisher (1954). Here, p-values are combined by Hartung's (1999) approach, which builds on the inverse normal method. It relies on the so-called probits $\tau_i = \Phi^{-1}(p_i)$, where p_i corresponds to the individual p-values. Linear combining the probits, i.e. $\sum_{i=1}^{N} \lambda_i \tau_i$, with $\lambda_1 \ldots, \lambda_N$, yields a normally distributed test statistic. Hartung (1999) accounts for a constant correlation r between these probits, which is estimated by $\hat{r}^* = \max(-\frac{1}{N-1}, \hat{r})$ with $\hat{r} = 1 - \frac{1}{N-1} \sum_{i=1}^{N} (\tau_i - \bar{\tau})^2$, where $\bar{\tau}$ represents the mean over the probits. It yields the test statistic (with equal weights $\lambda_i = 1$):

Modified Diebold-Mariano Test Results for m=ANN-WQ

j	$k=1$	$k=2$	$k=3$	$k=4$
Linear-WQ	0.0762*	0.0440**	0.0305**	0.0260**
Linear-BW	0.2042	0.1817	0.1607	0.1432
Linear-FW	0.0065***	0.0334**	0.0257**	0.0213**
LSTAR-WQ	0.1723	0.0769*	0.0586*	0.0467**
LSTAR-BW	0.1447	0.1076	0.0875*	0.0665*
LSTAR-FW	0.0353**	0.0204**	0.0168**	0.0217**
ESTAR-WQ	0.1457	0.0315**	0.0266**	0.0300**
ESTAR-BW	0.1302	0.0813*	0.0628*	0.0516*
ESTAR-FW	0.0254**	0.0434**	0.0256**	0.0216**
L2STAR-WQ	0.1183	0.0535*	0.0373**	0.0385**
L2STAR-BW	0.1636	0.1122	0.0831*	0.0659*
L2STAR-FW	0.0153**	0.0337**	0.0210**	0.0193**
ANN-BW	0.0001***	0.0034***	0.0077***	0.0219**
ANN-FW	0.0003***	0.0044***	0.0095***	0.0313**
AR(2)	0.2795	0.1666	0.1327	0.1071
ARIMA(1,1,0)	0.3419	0.1745	0.1381	0.1037
Mean	0.3210	0.1929	0.1588	0.1418

Table 3: The table reports p-values of the MDM test with H_0: $MSFE_{ANN-WQ} \geq MSFE_j$. */**/*** denote rejection of the null hypothesis at $\alpha = 10\%/5\%/1\%$, respectively.

$$Har = \frac{\sum_{i=1}^{N} \tau_i}{\sqrt{N + [N^2 - N]\left[\hat{r}^* + 0.2\sqrt{\frac{2}{N+1}}(1-\hat{r}^*)\right]}},$$

which is compared to critical values from the standard normal distribution. H_0 is rejected for too small values of the test statistic. The overall significance level is hence given by Φ(Har). Table 5 presents the combined p-values of the overall null hypothesis that the ANN-WQ is *not* the best forecasting model. As can be seen, the null hypothesis can be rejected at $\alpha = 5\%$ for all forecasting horizons. Hence, there is statistical evidence that the ANN-WQ serves as the best forecasting tool[14].

[14]Bonferroni-type tests, along the lines of Simes (1986), which are not reported here, also support this result.

Modified Diebold-Mariano Test Results for j=ANN-WQ

m	$k=1$	$k=2$	$k=3$	$k=4$
Linear-WQ	0.9238	0.9231	0.9695	0.9740
Linear-BW	0.7958	0.8924	0.8393	0.8568
Linear-FW	0.9935	0.9666	0.9743	0.9787
LSTAR-WQ	0.8277	0.9231	0.9414	0.9533
LSTAR-BW	0.8553	0.8924	0.9125	0.9335
LSTAR-FW	0.9647	0.9796	0.9832	0.9783
ESTAR-WQ	0.8543	0.9685	0.9734	0.9700
ESTAR-BW	0.8698	0.9187	0.9372	0.9484
ESTAR-FW	0.9746	0.9566	0.9744	0.9784
L2STAR-WQ	0.8817	0.9465	0.9627	0.9615
L2STAR-BW	0.8364	0.8878	0.9169	0.9341
L2STAR-FW	0.9847	0.9663	0.9790	0.9807
ANN-BW	0.9999	0.9966	0.9923	0.9781
ANN-FW	0.9997	0.9956	0.9905	0.9687
AR(2)	0.7205	0.8334	0.8673	0.8929
ARIMA(1,1,0)	0.6581	0.8255	0.8619	0.8963
Mean	0.6790	0.8071	0.8412	0.8582

Table 4: The table reports p-values of the MDM test with H_0: $MSFE_m \geq MSFE_{ANN-WQ}$.

Hartung Test Results

	$k=1$	$k=2$	$k=3$	$k=4$
p_{Har}	0.0001	0.0324	0.0317	0.0377

Table 5: The table reports combined p-values of the Hartung (1999) approach testing the overall null hypothesis that the ANN-WQ is *not* the superior forecasting model for forecasting horizons $k = 1, \ldots, 4$.

4 Discussion

4.1 Crisis Analysis

The forecasted sample from 2001:Q2-2012:Q4 includes crisis periods where the federal funds rate was stuck at the so-called zero lower bound (ZLB). In order to check the results' robustness, the forecasted sample is splitted here in *pre-crisis* and *post-crisis* periods with the latter starting in 2007:Q3. Table 6 reports the forecasting rankings for the two subperiods according to the RMSFEs for forecasting horizons $k = 1, \ldots, 4$ and the rank *differences* (*Pre-Post rank*). The insights from this analysis are threefold. First, the superior forecasting performance of the ANN-WQ model is robust with respect to the sample under investigation. Except for the one-quarter ahead forecasts in the pre-crisis sample, where the ANN-WQ model is ranked on place three[15], it is always ranked first irrespective of the crisis or non-crisis periods. It's dominance is also increasing in the forecasting horizon for both subsamples, which can be seen from the RMSFEs which are available upon request. Second, the ANN-FW model is the one that improves the most switching from pre- to post-crisis periods. Third, the models with the largest losses in terms of rank differences between the two periods are mainly linear models. That possibly reflects a higher degree of nonlinearity since the crisis, also due to the zero lower bound.

4.2 Data separation

As explained in Section 2.2, the configuration and estimation of the ANNs requires data separation into a training and a validation set. The latter serves two purposes. On the one hand, the mean squared error in the validation set is used to determine the number of hidden units in the network and the initial weights endogenously. On the other hand, it provides an early stopping criteria for the Levenberg-Marquardt algorithm. In the baseline framework, the first 80% of the "in-sample"[16] data assemble the training and the last 20% the validation set. There is no distinct rule on how to choose the data splitting percentages, but usually the validation set consists of 10-30% of the data. Obviously, choosing a different data base may lead to different model specifications. For this application, the 80/20% splitting performed best for determining the number of neurons and initial weights. Keeping this configuration fixed, the results are robust with respect to different splittings for the early stopping procedure. Table 7 presents the RMSFEs for the other splittings,

[15] The differences of the RMSFEs between the AR2, the Linear-BW and the ANN-WQ model are very small though.

[16] After holding out the last 50% of the total sample for pseudo-out-of-sample forecasts.

Forecasting Ranks Pre- and Post-Crisis

	$k=1$			$k=2$			$k=3$			$k=4$		
Model	Pre	Post	Diff.	Pre	Post	Diff.	Pre	Post	Diff.	Pre	Post	Diff.
Linear-WQ	5	14	-9	5	16	-11	3	16	-13	5	16	-11
Linear-BW	2	13	-11	3	11	-8	6	9	-3	7	9	-2
Linear-FW	10	17	-7	9	17	-8	7	17	-10	8	17	-9
LSTAR-WQ	12	6	6	13	4	9	14	4	10	13	3	10
LSTAR-BW	9	10	-1	12	10	2	15	10	5	16	10	6
LSTAR-FW	16	7	9	15	8	7	10	8	2	10	7	3
ESTAR-WQ	13	3	10	14	7	7	11	6	5	11	5	6
ESTAR-BW	8	11	-3	10	13	-3	16	13	3	15	12	3
ESTAR-FW	11	15	-4	8	14	-6	5	14	-9	3	14	-11
L2STAR-WQ	14	9	5	16	6	10	13	7	6	9	6	3
L2STAR-BW	4	12	-8	6	12	-6	9	12	-3	12	11	1
L2STAR-FW	15	16	-1	11	15	-4	8	15	-7	6	15	-9
ANN-WQ	3	1	2	1	1	0	1	1	0	1	1	0
ANN-BW	18	18	0	18	18	0	18	18	0	18	18	0
ANN-FW	17	5	12	17	5	12	17	3	14	17	2	15
AR(2)	1	8	-7	2	9	-7	4	11	-7	4	13	-9
ARIMA(1,1,0)	6	4	2	7	3	4	12	5	7	14	8	6
Mean	7	2	5	4	2	2	2	2	0	2	4	-2

Table 6: Forecasting ranks for pre- (2001:Q2-2007:Q2) and post-crisis (2007:Q3-2012:Q4) for the forecasting horizons $h = 1, \ldots, 4$ according to the root mean squared forecasting errors (RMSFEs). Diff. denotes the rank differences between the two periods (Pre-Post rank).

keeping the hidden units configuration based on the 80/20% separation constant. The RMSFEs of the ANN-WQ model are slightly larger with the different data splittings compared to the 80/20% benchmark. However, it still yields the best forecasting performance over all models[17]. Using the 90/10% splitting, the ANN-BW and the ANN-FW models can improve slightly over the benchmark case. However, it does not change their relative performance to the other models.

Forecasting Results (RMSFEs) for Different Data Splittings

Splitting		ANN-WQ	ANN-BW	ANN-FW
70/30%	$k=1$	0.3781	0.5571	0.4917
	$k=2$	0.7152	1.0734	0.8755
	$k=3$	1.0508	1.5189	1.3160
	$k=4$	1.3980	1.8343	1.6612
90/10%	$k=1$	0.3687	0.5050	0.4575
	$k=2$	0.7013	1.0267	0.8687
	$k=3$	1.0094	1.5143	1.2236
	$k=4$	1.3255	1.9086	1.5559

Table 7: Root mean squared forecasting errors (RMSFEs) for forecasting horizons $k = 1, \ldots, 4$ after expanding window regressions for different early stopping data splittings (training/ validation set). The initial estimation period is 1987:Q3-2000:Q2. The structure of the WQ-/ BW-/ FW-ANN consists of 4, 1 and 2 hidden units, respectively, as determined by the 80/20% splitting of the benchmark case.

4.3 Linearity tests

To further investigate if U.S. monetary policy is linear or nonlinear, this section performs different linearity tests on the whole sample from 1987:Q3-2012:Q4. The tests taken under consideration are the Luukkonen et al. (1988), the Teräsvirta (1994) sequential and Escribano and Jorda (1999) test. All of them test for linearity against STR alternatives by testing $\gamma = 0$. Under the null hypothesis, the parameters c and β are not identified. Hence the transition function $G(\gamma, c, s_t)$ needs to be replaced by a Taylor series expansion in order to get the null distribution of the test statistic. Since this expansion depends on the specific form of $G(\cdot)$, it is possible to discriminate between different transition functions. Tables 17-19 in the Appendix report the test results for STR-WQ, -BW and -FW, respectively. The null hypothesis of linearity is rejected for the WQ and the FW specification, but cannot be rejected for the BW version. Hence, the results indicate that

[17] Except for the 70/30% case, where it is beaten by the univariate models for $k = 1$ and $k = 2$.

using now- or forecasts as explanatory variables already introduces nonlinearity. It may also explain why the Linear-BW model outperforms all other BW models and why the ANN-BW model is the worst in the forecasting exercise. The ANN can only be superior if there is enough nonlinearity, which seems not to be the case when using BW-inputs. For the WQ and the FW version, the Teräsvirta (1994) test suggests the LSTR model, while the Escribano-Jorda (1999) test recommends the ESTR model. The fact, that linearity is rejected for these input versions is in line with the finding that all WQ-STR and FW-STR models produced better forecasts than their linear counterparts. The differences in RMSFEs between the specific transition functions are small, though.

5 Conclusion

Using quarterly U.S. real-time data from 1987:Q3-2012:Q4, the paper shows that the artificial neural network is flexible enough to predict the federal funds rate better than linear and nonlinear Taylor rules as well as univariate processes. Specifically, it is the "within-quarter" specification with nowcasts of inflation and the output gap and two lags of the federal funds rate as explanatory variables that yields the smallest root mean squared forecast errors over all forecasting horizons (one- to four-quarters ahead). The result is robust with respect to different time periods indicating that the artificial neural network is a useful forecasting tool for normal as well as crisis times. It is also robust with respect to different data splittings in the estimation phase. Linearity tests indicate that using now- and forecasts of inflation and the output gap introduces nonlinearity, while linearity cannot be rejected with backcasts of the explanatory variables.

The paper at hand has shown the potential of artificial neural networks as a forecasting tool for U.S. monetary policy. Future work could include more explanatory variables as e.g. asset purchases in crisis times or financial stability indicators to check whether the forecasts can be improved. A similar analysis could be undertaken for monetary policy in the euro area as well. Generally, the results also suggest the worthiness of real-time forecasts of the explanatory variables in the reaction function. If the Fed aims at explicit forward guidance, it might be easier for the market if the Fed publishes its current Greenbook forecasts without a delay.

References

S. Aras and İ. D. Kocakoç. A new model selection strategy in time series forecasting with artificial neural networks: IHTS. *Neurocomputing*, 174:974–987, 2016.

B. S. Bernanke and J. Boivin. Monetary policy in a data-rich environment. *Journal of Monetary Economics*, 50(3):525–546, 2003.

V. Castro. Can central banks' monetary policy be described by a linear (augmented) Taylor rule or by a nonlinear rule? *Journal of Financial Stability*, 7(4):228–246, 2011.

R. Clarida, J. Galí, and M. Gertler. Monetary policy rules in practice: Some international evidence. *European Economic Review*, 42(6):1033–1067, 1998.

R. Clarida, J. Galí, and M. Gertler. Monetary policy rules and macroeconomic stability: Evidence and some theory. *The Quarterly Journal of Economics*, 115(1):147–180, 2000.

T. E. Clark and M. W. McCracken. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1):85–110, 2001.

T. E. Clark and M. W. McCracken. Nested forecast model comparisons: A new approach to testing equal accuracy. *Manuscript, Federal Reserve Banks of Cleveland and St. Louis*, 2011.

T. E. Clark and M. W. McCracken. Advances in forecast evaluation. In *Handbook of Economic Forecasting*, volume 2, pages 1107–1201. Elsevier, 2013.

A. Cukierman and A. Muscatelli. Nonlinear Taylor rules and asymmetric preferences in central banking: Evidence from the United Kingdom and the United States. *The BE Journal of Macroeconomics*, 8(1), 2008.

F. X. Diebold. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics*, 33(1):1–23, 2015.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13:253–263, 1995.

J. J. Dolado, R. M.-D. Pedrero, and F. J. Ruge-Murcia. Nonlinear monetary policy rules: Some new evidence for the US. *Studies in Nonlinear Dynamics & Econometrics*, 8(3), 2004.

J. J. Dolado, R. María-Dolores, and M. Naveira. Are monetary-policy reaction functions asymmetric?: The role of nonlinearity in the Phillips curve. *European Economic Review*, 49(2):485–503, 2005.

M. Draghi. The Outlook for the Euro Area Economy, 2018. Frankfurt European Banking Congress, Frankfurt am Main, 16 November.

A. Escribano and O. Jorda. Improved testing and specification of smooth transition regression models. In *Nonlinear time series analysis of economic and financial data*, pages 289–319. Springer, 1999.

R. A. Fisher. *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh, 12th edition, 1954.

S. Gonzalez. Neural networks for macroeconomic forecasting: A complementary approach to linear regression models. *Working Paper, Department of Finance Canada*, 2000.

M. Goodfriend. Inflation targeting in the United States? In *The Inflation-Targeting Debate*, pages 311–352. University of Chicago Press, 2004.

J. Hartung. A note on combining dependent tests of significance. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 41(7):849–855, 1999.

D. Harvey, S. Leybourne, and P. Newbold. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291, 1997.

S. S. Haykin. Neural networks: A comprehensive foundation, 1999.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

D. H. Kim, D. R. Osborn, and M. Sensier. Nonlinearity in the Fed's monetary policy rule. *Journal of Applied Econometrics*, 20(5):621–639, 2005.

Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

R. Luukkonen, P. Saikkonen, and T. Teräsvirta. Testing linearity against smooth transition autoregressive models. *Biometrika*, 75(3):491–499, 1988.

A. G. Malliaris and M. Malliaris. Modeling federal funds rates: A comparison of four methodologies. *Neural Computing and Applications*, 18(1):37–44, 2009.

M. W. McCracken. Asymptotics for out of sample tests of causality. *Manuscript, Louisiana State University*, 1999.

R. A. Nobay and D. A. Peel. Optimal discretionary monetary policy in a model of asymmetric central bank preferences. *The Economic Journal*, 113(489):657–665, 2003.

A. Orphanides. Monetary policy rules based on real-time data. *American Economic Review*, 91(4):964–985, 2001.

A. Orphanides and V. Wieland. Economic projections and rules-of-thumb for monetary policy. *CFS Working Paper No. 2008/16*, 2008.

K. Petersen. Does the Federal Reserve follow a non-linear Taylor rule? *Economics Working Papers. 200737*, 2007.

T. Qin and W. Enders. In-sample and out-of-sample properties of linear and nonlinear Taylor rules. *Journal of Macroeconomics*, 30(1):428–443, 2008.

F. J. Ruge-Murcia. The inflation bias when the central bank targets the natural rate of unemployment. *European Economic Review*, 48(1):91–107, 2004.

E. Schaling. The nonlinear Phillips curve and inflation forecast targeting: Symmetric versus asymmetric monetary policy rules. *Journal of Money, Credit and Banking*, 36 (3):361–386, 2004.

R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

L. Svensson. Inflation forecast targeting: Implementing and monitoring inflation targets. *European Economic Review*, 41(6):1111–1146, 1997.

J. B. Taylor. Discretion versus policy rules in practice. In *Carnegie-Rochester conference series on public policy*, volume 39, pages 195–214, 1993.

T. Teräsvirta. Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, 89(425):208–218, 1994.

T. Teräsvirta, D. Van Dijk, and M. C. Medeiros. Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: A reexamination. *International Journal of Forecasting*, 21(4):755–774, 2005.

T. Teräsvirta, D. Tjøstheim, and C. W. J. Granger. Modelling nonlinear economic time series, 2010.

K. D. West. Forecast evaluation. In *Handbook of Economic Forecasting*, volume 1, pages 99–134. Elsevier, 2006.

J. Yellen. The economic outlook and the conduct of monetary policy, 2017. Remarks at Stanford Institute for Economic Policy Research, Stanford University, 19 January.

Appendix

List of Models

1) Linear-WQ:	$i_t = (1-\rho)(\alpha + \theta_\pi \pi_{t\|t} + \beta_y y_{t\|t}) + \rho_1 i_{t-1} + \rho_2 i_{t-1} + \varepsilon_t$
2) Linear-BW:	$i_t = (1-\rho)(\alpha + \theta_\pi \pi_{t-1\|t} + \beta_y y_{t-1\|t}) + \rho_1 i_{t-1} + \rho_2 i_{t-1} + \varepsilon_t$
3) Linear-FW:	$i_t = (1-\rho)(\alpha + \theta_\pi \pi_{t+1\|t} + \beta_y y_{t+1\|t}) + \rho_1 i_{t-1} + \rho_2 i_{t-1} + \varepsilon_t$
4) LSTR-WQ:	$i_t = \alpha_0 + \alpha_1 \pi_{t\|t} + \alpha_2 y_{t\|t} + \alpha_3 i_{t-1} + \alpha_4 i_{t-2} +$ $\{1 + exp[-\gamma(s_t - c)]\}^{-1}(\beta_0 + \beta_1 \pi_{t\|t} + \beta_2 y_{t\|t} + \beta_3 i_{t-1} + \beta_4 i_{t-2}) + \varepsilon_t$
5) LSTR-BW:	$i_t = \alpha_0 + \alpha_1 \pi_{t-1\|t} + \alpha_2 y_{t-1\|t} + \alpha_3 i_{t-1} + \alpha_4 i_{t-2} +$ $\{1 + exp[-\gamma(s_t - c)]\}^{-1}(\beta_0 + \beta_1 \pi_{t-1\|t} + \beta_2 y_{t-1\|t} + \beta_3 i_{t-1} + \beta_4 i_{t-2}) + \varepsilon_t$
6) LSTR-FW:	$i_t = \alpha_0 + \alpha_1 \pi_{t+1\|t} + \alpha_2 y_{t+1\|t} + \alpha_3 i_{t-1} + \alpha_4 i_{t-2} +$ $\{1 + exp[-\gamma(s_t - c)]\}^{-1}(\beta_0 + \beta_1 \pi_{t+1\|t} + \beta_2 y_{t+1\|t} + \beta_3 i_{t-1} + \beta_4 i_{t-2}) + \varepsilon_t$
7) ESTR-WQ:	$i_t = \alpha_0 + \alpha_1 \pi_{t\|t} + \alpha_2 y_{t\|t} + \alpha_3 i_{t-1} + \alpha_4 i_{t-2} +$ $\{1 - exp[-\gamma(s_t - c)^2]\}(\beta_0 + \beta_1 \pi_{t\|t} + \beta_2 y_{t\|t} + \beta_3 i_{t-1} + \beta_4 i_{t-2}) + \varepsilon_t$
8) ESTR-BW:	$i_t = \alpha_0 + \alpha_1 \pi_{t-1\|t} + \alpha_2 y_{t-1\|t} + \alpha_3 i_{t-1} + \alpha_4 i_{t-2} +$ $\{1 - exp[-\gamma(s_t - c)^2]\}(\beta_0 + \beta_1 \pi_{t-1\|t} + \beta_2 y_{t-1\|t} + \beta_3 i_{t-1} + \beta_4 i_{t-2}) + \varepsilon_t$
9) ESTR-FW:	$i_t = \alpha_0 + \alpha_1 \pi_{t+1\|t} + \alpha_2 y_{t+1\|t} + \alpha_3 i_{t-1} + \alpha_4 i_{t-2} +$ $\{1 - exp[-\gamma(s_t - c)^2]\}(\beta_0 + \beta_1 \pi_{t+1\|t} + \beta_2 y_{t+1\|t} + \beta_3 i_{t-1} + \beta_4 i_{t-2}) + \varepsilon_t$
10) L2STR-WQ:	$i_t = \alpha_0 + \alpha_1 \pi_{t\|t} + \alpha_2 y_{t\|t} + \alpha_3 i_{t-1} + \alpha_4 i_{t-2} +$ $\{1 + exp[-\gamma(s_t - c_1)(s_t - c_2)]\}^{-1}(\beta_0 + \beta_1 \pi_{t\|t} + \beta_2 y_{t\|t} + \beta_3 i_{t-1} + \beta_4 i_{t-2}) + \varepsilon_t$
11) L2STR-BW:	$i_t = \alpha_0 + \alpha_1 \pi_{t-1\|t} + \alpha_2 y_{t-1\|t} + \alpha_3 i_{t-1} + \alpha_4 i_{t-2} +$ $\{1 + exp[-\gamma(s_t - c_1)(s_t - c_2)]\}^{-1}(\beta_0 + \beta_1 \pi_{t-1\|t} + \beta_2 y_{t-1\|t} + \beta_3 i_{t-1} + \beta_4 i_{t-2}) + \varepsilon_t$
12) L2STR-FW:	$i_t = \alpha_0 + \alpha_1 \pi_{t+1\|t} + \alpha_2 y_{t+1\|t} + \alpha_3 i_{t-1} + \alpha_4 i_{t-2} +$ $\{1 + exp[-\gamma(s_t - c_1)(s_t - c_2)]\}^{-1}(\beta_0 + \beta_1 \pi_{t+1\|t} + \beta_2 y_{t+1\|t} + \beta_3 i_{t-1} + \beta_4 i_{t-2}) + \varepsilon_t$
13) ANN-WQ:	$i_t = \alpha_0 + \sum_{j=1}^{q} \gamma_j \cdot G(\alpha_j + \beta_{1j} \pi_{t\|t} + \beta_{2j} y_{t\|t} + \beta_{3j} i_{t-1} + \beta_{4j} i_{t-2}) + \varepsilon_t$
14) ANN-BW:	$i_t = \alpha_0 + \sum_{j=1}^{q} \gamma_j \cdot G(\alpha_j + \beta_{1j} \pi_{t-1\|t} + \beta_{2j} y_{t-1\|t} + \beta_{3j} i_{t-1} + \beta_{4j} i_{t-2}) + \varepsilon_t$
15) ANN-FW:	$i_t = \alpha_0 + \sum_{j=1}^{q} \gamma_j \cdot G(\alpha_j + \beta_{1j} \pi_{t+1\|t} + \beta_{2j} y_{t+1\|t} + \beta_{3j} i_{t-1} + \beta_{4j} i_{t-2}) + \varepsilon_t$
16) AR(2):	$i_t = \alpha_0 + \alpha_1 i_{t-1} + \alpha_2 i_{t-2} + \varepsilon_t$
17) ARIMA(1,1,0):	$\Delta i_t = \alpha_0 + \alpha_1 \Delta i_{t-1} + \varepsilon_t$
18) Mean:	Equally weighted average over models 1-17

Table 8: Summary of models used in the forecasting performance comparison. The threshold variable s_t of the STR models is either inflation π or the output gap y. In the ANN specifications, $G(\cdot)$ denotes the hyperbolic tangent transfer function. The choice of q is explained in section 2.2.

ADF Test Results for Loss-Differentials D_{mjt} $(k = 1)$

m/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1		0.002	0.000	0.000	0.198	0.000	0.002	0.058	0.000	0.000	0.011	0.000	0.000	0.047	0.371	0.000	0.000	0.131
2			0.001	0.008	0.012	0.000	0.045	0.000	0.000	0.000	0.000	0.000	0.000	0.032	0.002	0.000	0.000	0.000
3				0.062	0.001	0.018	0.031	0.006	0.000	0.001	0.008	0.000	0.184	0.156	0.052	0.021	0.022	0.040
4					0.058	0.000	0.007	0.000	0.000	0.000	0.022	0.000	0.000	0.043	0.193	0.000	0.000	0.000
5						0.000	0.031	0.095	0.000	0.000	0.160	0.000	0.000	0.037	0.010	0.000	0.000	0.010
6							0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.101	0.016	0.000	0.000	0.000
7								0.032	0.000	0.000	0.064	0.000	0.000	0.064	0.002	0.000	0.000	0.000
8									0.000	0.000	0.000	0.000	0.000	0.040	0.002	0.000	0.000	0.000
9										0.000	0.000	0.001	0.000	0.079	0.003	0.000	0.000	0.000
10											0.000	0.000	0.000	0.009	0.311	0.000	0.000	0.000
11												0.000	0.000	0.037	0.002	0.000	0.000	0.001
12													0.148	0.068	0.000	0.000	0.000	0.000
13														0.174	0.001	0.000	0.000	0.000
14															0.035	0.032	0.012	0.116
15																0.006	0.001	0.004
16																	0.000	0.000
17																		0.000
18																		

Table 9: p-values of the Augmented Dickey Fuller (ADF) test with unit root null hypothesis for the loss differentials $D_{mjt} = (i_t - f_t^{(m)})^2 - (i_t - f_t^{(j)})^2$ and forecasting horizon $k = 1$. The lag length was determined by the Bayesian information criterion (BIC). See Table 8 for the model definitions.

ADF Test Results for Loss-Differentials D_{mjt} $(k = 2)$

m/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1		0.002	0.059	0.025	0.008	0.002	0.135	0.024	0.000	0.000	0.019	0.000	0.002	0.052	0.043	0.000	0.000	0.166
2			0.022	0.000	0.032	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.023	0.003	0.000	0.000	0.000
3				0.103	0.040	0.127	0.120	0.055	0.036	0.010	0.048	0.001	0.110	0.133	0.178	0.006	0.000	0.042
4					0.001	0.000	0.000	0.002	0.000	0.000	0.002	0.000	0.000	0.066	0.009	0.000	0.000	0.000
5						0.001	0.041	0.000	0.000	0.000	0.000	0.000	0.001	0.026	0.001	0.000	0.000	0.003
6							0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.038	0.012	0.000	0.000	0.000
7								0.002	0.000	0.000	0.001	0.000	0.000	0.040	0.005	0.000	0.000	0.000
8									0.000	0.000	0.000	0.000	0.000	0.030	0.002	0.000	0.000	0.000
9										0.000	0.000	0.000	0.211	0.056	0.008	0.000	0.000	0.000
10											0.000	0.000	0.000	0.017	0.000	0.000	0.000	0.000
11												0.000	0.000	0.031	0.002	0.000	0.000	0.000
12													0.000	0.051	0.002	0.000	0.000	0.000
13														0.066	0.010	0.000	0.000	0.000
14															0.004	0.024	0.010	0.053
15																0.000	0.000	0.017
16																	0.000	0.000
17																		0.000
18																		

Table 10: p-values of the Augmented Dickey Fuller (ADF) test with unit root null hypothesis for the loss differentials $D_{mjt} = (i_t - f_t^{(m)})^2 - (i_t - f_t^{(j)})^2$ and forecasting horizon $k = 2$. The lag length was determined by the Bayesian information criterion (BIC). See Table 8 for the model definitions.

ADF Test Results for Loss-Differentials D_{mjt} ($k = 3$)

m/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1		0.038	0.096	0.048	0.092	0.034	0.034	0.081	0.067	0.000	0.046	0.008	0.038	0.006	0.026	0.004	0.000	0.247
2			0.084	0.002	0.002	0.004	0.001	0.000	0.000	0.000	0.000	0.000	0.002	0.036	0.001	0.000	0.000	0.000
3				0.086	0.148	0.074	0.037	0.117	0.093	0.027	0.101	0.018	0.283	0.017	0.589	0.027	0.001	0.086
4					0.043	0.000	0.000	0.022	0.000	0.000	0.014	0.000	0.001	0.081	0.000	0.002	0.000	0.002
5						0.025	0.019	0.004	0.073	0.001	0.009	0.008	0.048	0.042	0.000	0.000	0.000	0.053
6							0.000	0.011	0.002	0.000	0.005	0.000	0.001	0.013	0.005	0.006	0.000	0.006
7								0.016	0.000	0.000	0.002	0.001	0.057	0.110	0.001	0.006	0.000	0.000
8									0.023	0.000	0.005	0.044	0.003	0.043	0.000	0.000	0.000	0.004
9										0.000	0.000	0.000	0.005	0.012	0.008	0.001	0.000	0.000
10											0.000	0.001	0.000	0.052	0.000	0.005	0.000	0.000
11												0.000	0.002	0.042	0.000	0.000	0.000	0.001
12													0.003	0.034	0.477	0.004	0.000	0.001
13														0.021	0.000	0.002	0.000	0.006
14															0.007	0.045	0.032	0.072
15																0.000	0.000	0.003
16																	0.000	0.000
17																		0.000
18																		

Table 11: p-values of the Augmented Dickey Fuller (ADF) test with unit root null hypothesis for the loss differentials $D_{mjt} = (i_t - f_t^{(m)})^2 - (i_t - f_t^{(j)})^2$ and forecasting horizon $k = 3$. The lag length was determined by the Bayesian information criterion (BIC). See Table 8 for the model definitions.

ADF Test Results for Loss-Differentials D_{mjt} ($k = 4$)

m/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1		0.056	0.066	0.045	0.130	0.075	0.060	0.092	0.036	0.006	0.054	0.021	0.006	0.063	0.006	0.028	0.001	0.114
2			0.071	0.003	0.003	0.011	0.002	0.000	0.000	0.000	0.002	0.000	0.000	0.057	0.001	0.002	0.000	0.001
3				0.094	0.138	0.045	0.027	0.113	0.030	0.036	0.009	0.009	0.065	0.008	0.016	0.004	0.006	0.124
4					0.105	0.003	0.001	0.035	0.001	0.000	0.028	0.001	0.000	0.112	0.000	0.036	0.000	0.011
5						0.067	0.067	0.021	0.191	0.014	0.066	0.260	0.013	0.069	0.000	0.020	0.001	0.106
6							0.000	0.025	0.005	0.000	0.017	0.001	0.000	0.010	0.000	0.032	0.001	0.024
7								0.016	0.124	0.043	0.010	0.105	0.010	0.137	0.000	0.043	0.000	0.003
8									0.004	0.001	0.000	0.110	0.003	0.062	0.000	0.000	0.000	0.013
9										0.001	0.001	0.000	0.002	0.008	0.001	0.057	0.001	0.070
10											0.001	0.003	0.000	0.011	0.000	0.021	0.000	0.000
11												0.115	0.003	0.061	0.000	0.001	0.000	0.005
12													0.003	0.013	0.002	0.020	0.001	0.002
13														0.014	0.001	0.014	0.001	0.000
14															0.073	0.075	0.077	0.103
15																0.000	0.002	0.000
16																	0.000	0.019
17																		0.000
18																		

Table 12: p-values of the Augmented Dickey Fuller (ADF) test with unit root null hypothesis for the loss differentials $D_{mjt} = (i_t - f_t^{(m)})^2 - (i_t - f_t^{(j)})^2$ and forecasting horizon $k = 4$. The lag length was determined by the Bayesian information criterion (BIC). See Table 8 for the model definitions.

Modified Diebold-Mariano Test Results ($k = 1$)

m/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	-	0.92	0.02	0.84	0.84	0.50	0.86	0.86	0.36	0.52	0.93	0.13	0.92	0.00	0.01	0.86	0.78	0.99
2	0.08	-	0.02	0.40	0.38	0.21	0.43	0.24	0.15	0.18	0.30	0.06	0.80	0.00	0.00	0.72	0.65	0.85
3	0.98	0.98	-	0.97	0.97	0.94	0.97	0.98	0.93	0.87	0.98	0.68	0.99	0.00	0.08	0.98	0.94	1.00
4	0.16	0.60	0.03	-	0.53	0.14	0.57	0.50	0.18	0.27	0.54	0.04	0.83	0.00	0.00	0.69	0.65	0.92
5	0.16	0.62	0.03	0.47	-	0.24	0.50	0.39	0.18	0.24	0.49	0.08	0.86	0.00	0.00	0.82	0.69	0.91
6	0.50	0.79	0.06	0.86	0.76	-	0.90	0.76	0.35	0.52	0.78	0.11	0.96	0.00	0.01	0.84	0.78	0.99
7	0.14	0.57	0.03	0.43	0.50	0.10	-	0.45	0.17	0.23	0.50	0.03	0.85	0.00	0.00	0.70	0.65	0.96
8	0.14	0.76	0.02	0.50	0.61	0.24	0.55	-	0.18	0.26	0.67	0.06	0.87	0.00	0.00	0.82	0.70	0.98
9	0.64	0.85	0.07	0.82	0.82	0.65	0.83	0.82	-	0.61	0.84	0.14	0.97	0.00	0.02	0.90	0.85	0.96
10	0.48	0.82	0.13	0.73	0.76	0.48	0.77	0.74	0.39	-	0.50	0.03	0.88	0.00	0.01	0.70	0.65	0.92
11	0.07	0.70	0.02	0.46	0.51	0.22	0.50	0.33	0.16	0.22	-	0.06	0.84	0.00	0.00	0.82	0.70	0.96
12	0.87	0.94	0.32	0.96	0.92	0.89	0.97	0.94	0.86	0.84	0.95	-	0.98	0.00	0.05	0.90	0.85	0.99
13	0.08	0.20	0.01	0.17	0.14	0.04	0.15	0.13	0.03	0.12	0.16	0.02	-	0.00	0.00	0.28	0.34	0.32
14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-	0.98	1.00	1.00	1.00
15	0.99	1.00	0.92	1.00	1.00	0.99	1.00	1.00	0.98	0.99	1.00	0.95	1.00	0.02	-	1.00	0.99	1.00
16	0.14	0.28	0.02	0.31	0.18	0.16	0.30	0.18	0.10	0.15	0.25	0.05	0.72	0.00	0.00	-	0.53	0.57
17	0.22	0.35	0.06	0.35	0.31	0.22	0.35	0.30	0.15	0.21	0.33	0.09	0.66	0.00	0.01	0.47	-	0.53
18	0.01	0.15	0.00	0.08	0.09	0.01	0.04	0.02	0.04	0.08	0.04	0.01	0.68	0.00	0.00	0.43	0.47	-

Table 13: p-values of the Modified Diebold Mariano (MDM) test of Harvey et al. (1997) with the null hypothesis $MSFE_m \geq MSFE_j$ and forecasting horizon $k = 1$. See Table 8 for the model definitions.

Diebold-Mariano Test Results ($k = 2$)

m/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	-	0.93	0.09	0.83	0.76	0.73	0.77	0.81	0.67	0.72	0.89	0.49	0.92	0.01	0.02	0.90	0.86	0.96
2	0.24	-	0.09	0.83	0.76	0.73	0.77	0.81	0.67	0.72	0.89	0.49	0.89	0.00	0.01	0.90	0.86	0.79
3	0.91	0.93	-	0.90	0.87	0.89	0.87	0.88	0.89	0.86	0.91	0.88	0.97	0.02	0.35	0.92	0.92	0.95
4	0.17	0.65	0.10	-	0.44	0.26	0.43	0.43	0.23	0.35	0.54	0.17	0.92	0.01	0.02	0.67	0.67	0.79
5	0.24	0.78	0.13	0.56	-	0.44	0.52	0.51	0.36	0.46	0.69	0.27	0.89	0.00	0.01	0.81	0.71	0.92
6	0.27	0.73	0.11	0.74	0.56	-	0.66	0.57	0.35	0.54	0.67	0.23	0.98	0.01	0.04	0.75	0.75	0.95
7	0.23	0.67	0.13	0.57	0.48	0.34	-	0.47	0.25	0.40	0.58	0.20	0.97	0.01	0.03	0.72	0.73	0.96
8	0.19	0.91	0.12	0.57	0.49	0.43	0.53	-	0.33	0.45	0.96	0.24	0.92	0.01	0.02	0.90	0.75	0.97
9	0.33	0.83	0.11	0.77	0.64	0.65	0.75	0.67	-	0.63	0.76	0.23	0.96	0.01	0.09	0.83	0.82	0.95
10	0.28	0.72	0.14	0.65	0.54	0.46	0.60	0.55	0.37	-	0.65	0.26	0.95	0.01	0.02	0.74	0.74	0.93
11	0.11	0.79	0.09	0.46	0.31	0.33	0.42	0.04	0.24	0.35	-	0.17	0.89	0.01	0.01	0.76	0.67	0.95
12	0.51	0.88	0.12	0.83	0.73	0.77	0.80	0.76	0.77	0.74	0.83	-	0.97	0.01	0.12	0.87	0.87	0.96
13	0.04	0.18	0.03	0.08	0.11	0.02	0.03	0.08	0.04	0.05	0.11	0.03	-	0.00	0.00	0.17	0.17	0.19
14	0.99	0.99	0.98	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	-	0.96	0.99	0.99	1.00
15	0.89	0.99	0.65	0.98	0.99	0.96	0.97	0.98	0.91	0.98	0.99	0.88	1.00	0.04	-	0.98	0.98	1.00
16	0.10	0.41	0.08	0.33	0.19	0.25	0.28	0.10	0.17	0.26	0.24	0.13	0.83	0.01	0.02	-	0.52	0.71
17	0.14	0.43	0.08	0.33	0.29	0.25	0.27	0.25	0.18	0.26	0.33	0.13	0.83	0.01	0.02	0.48	-	0.65
18	0.04	0.21	0.05	0.08	0.10	0.05	0.04	0.03	0.05	0.07	0.05	0.04	0.81	0.00	0.00	0.29	0.35	-

Table 14: p-values of the Modified Diebold Mariano (MDM) test of Harvey et al. (1997) with the null hypothesis $MSFE_m \geq MSFE_j$ and forecasting horizon $k = 2$. See Table 8 for the model definitions.

Diebold-Mariano Test Results ($k = 3$)

m/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	-	0.88	0.10	0.84	0.70	0.82	0.81	0.68	0.77	0.80	0.80	0.61	0.97	0.01	0.26	0.83	0.78	0.96
2	0.12	-	0.10	0.44	0.22	0.41	0.44	0.04	0.28	0.39	0.13	0.19	0.84	0.01	0.02	0.45	0.45	0.81
3	0.90	0.90	-	0.89	0.82	0.91	0.87	0.81	0.88	0.87	0.86	0.87	0.97	0.02	0.51	0.88	0.86	0.94
4	0.16	0.56	0.11	-	0.37	0.41	0.49	0.28	0.27	0.40	0.38	0.18	0.94	0.02	0.04	0.52	0.50	0.89
5	0.30	0.78	0.18	0.63	-	0.58	0.62	0.37	0.48	0.60	0.61	0.37	0.91	0.01	0.05	0.73	0.62	0.90
6	0.18	0.59	0.09	0.59	0.42	-	0.56	0.36	0.30	0.50	0.45	0.16	0.98	0.02	0.07	0.56	0.54	0.85
7	0.19	0.56	0.13	0.51	0.38	0.44	-	0.30	0.24	0.43	0.40	0.19	0.97	0.02	0.07	0.53	0.51	0.90
8	0.32	0.96	0.19	0.72	0.63	0.64	0.70	-	0.53	0.68	0.94	0.40	0.94	0.02	0.07	0.87	0.70	0.98
9	0.23	0.72	0.12	0.73	0.52	0.70	0.76	0.47	-	0.67	0.58	0.21	0.97	0.02	0.15	0.68	0.66	0.94
10	0.20	0.61	0.13	0.60	0.40	0.50	0.57	0.32	0.33	-	0.43	0.22	0.96	0.02	0.04	0.57	0.55	0.93
11	0.20	0.87	0.14	0.62	0.39	0.55	0.60	0.06	0.42	0.57	-	0.30	0.92	0.01	0.04	0.71	0.60	0.96
12	0.39	0.81	0.13	0.82	0.63	0.84	0.81	0.60	0.79	0.78	0.70	-	0.98	0.02	0.23	0.76	0.74	0.94
13	0.03	0.16	0.03	0.06	0.09	0.02	0.03	0.06	0.03	0.04	0.08	0.02	-	0.01	0.01	0.13	0.14	0.16
14	0.99	0.99	0.98	0.98	0.99	0.98	0.98	0.98	0.98	0.98	0.99	0.98	0.99	-	0.95	0.99	0.98	0.99
15	0.74	0.98	0.49	0.96	0.95	0.93	0.93	0.93	0.85	0.96	0.96	0.77	0.99	0.05	-	0.95	0.91	0.99
16	0.17	0.55	0.12	0.48	0.27	0.44	0.47	0.13	0.32	0.43	0.29	0.24	0.87	0.01	0.05	-	0.47	0.80
17	0.22	0.55	0.14	0.50	0.38	0.46	0.49	0.30	0.34	0.45	0.40	0.26	0.86	0.02	0.09	0.53	-	0.75
18	0.04	0.19	0.06	0.11	0.10	0.15	0.10	0.02	0.06	0.07	0.04	0.06	0.84	0.01	0.01	0.20	0.25	-

Table 15: p-values of the Modified Diebold Mariano (MDM) test of Harvey et al. (1997) with the null hypothesis $MSFE_m \geq MSFE_j$ and forecasting horizon $k = 3$. See Table 8 for the model definitions.

Diebold-Mariano Test Results ($k = 4$)

m/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	-	0.84	0.14	0.85	0.59	0.87	0.85	0.59	0.72	0.87	0.71	0.62	0.97	0.04	0.45	0.74	0.66	0.95
2	0.16	-	0.15	0.53	0.18	0.53	0.55	0.03	0.28	0.55	0.11	0.25	0.86	0.04	0.12	0.33	0.34	0.84
3	0.86	0.85	-	0.88	0.71	0.91	0.88	0.72	0.84	0.89	0.79	0.83	0.98	0.04	0.61	0.80	0.75	0.93
4	0.15	0.47	0.12	-	0.28	0.53	0.58	0.21	0.20	0.57	0.29	0.17	0.95	0.04	0.10	0.39	0.33	0.85
5	0.41	0.82	0.29	0.72	-	0.71	0.73	0.46	0.53	0.74	0.66	0.48	0.93	0.05	0.35	0.70	0.57	0.91
6	0.13	0.47	0.09	0.47	0.29	-	0.53	0.25	0.16	0.51	0.31	0.11	0.98	0.04	0.13	0.40	0.35	0.76
7	0.15	0.45	0.12	0.42	0.27	0.47	-	0.21	0.15	0.46	0.28	0.15	0.97	0.05	0.14	0.36	0.29	0.80
8	0.41	0.97	0.28	0.79	0.54	0.75	0.79	-	0.56	0.81	0.94	0.49	0.95	0.05	0.34	0.79	0.61	0.98
9	0.28	0.72	0.16	0.80	0.47	0.84	0.85	0.44	-	0.84	0.54	0.27	0.98	0.05	0.34	0.61	0.55	0.94
10	0.13	0.45	0.11	0.43	0.26	0.49	0.54	0.19	0.16	-	0.26	0.15	0.96	0.04	0.11	0.36	0.31	0.88
11	0.29	0.89	0.21	0.71	0.34	0.69	0.72	0.06	0.46	0.74	-	0.40	0.93	0.04	0.25	0.63	0.52	0.96
12	0.38	0.75	0.17	0.83	0.52	0.89	0.85	0.51	0.73	0.85	0.60	-	0.98	0.05	0.40	0.65	0.60	0.93
13	0.03	0.14	0.02	0.05	0.07	0.02	0.03	0.05	0.02	0.04	0.07	0.02	-	0.02	0.03	0.11	0.10	0.14
14	0.96	0.96	0.96	0.96	0.95	0.96	0.95	0.95	0.95	0.96	0.96	0.95	0.98	-	0.92	0.95	0.94	0.97
15	0.55	0.88	0.39	0.90	0.65	0.87	0.86	0.66	0.66	0.89	0.75	0.60	0.97	0.08	-	0.76	0.70	0.97
16	0.26	0.67	0.20	0.61	0.30	0.60	0.64	0.21	0.39	0.64	0.37	0.35	0.89	0.05	0.24	-	0.43	0.86
17	0.34	0.66	0.25	0.67	0.43	0.65	0.71	0.39	0.45	0.69	0.48	0.40	0.90	0.06	0.30	0.57	-	0.84
18	0.05	0.16	0.07	0.15	0.09	0.24	0.20	0.02	0.06	0.12	0.04	0.07	0.86	0.03	0.03	0.14	0.16	-

Table 16: p-values of the Modified Diebold Mariano (MDM) test of Harvey et al. (1997) with the null hypothesis $MSFE_m \geq MSFE_j$ and forecasting horizon $k = 4$. See Table 8 for the model definitions.

Linearity Tests for STR-WQ

Taylor series alternatives: $b_0 + b_1 \cdot s[+ b_2 \cdot s^2 + b_3 \cdot s^3 + b_4 \cdot s^4]$
Threshold variable s: Y

Luukkonen, Saikkonen, and Teräsvirta (1988) Linearity Tests

Null Hypothesis	F-statistic	d.f.	p-value
$H_0^{(4)}$: b_1=b_2=b_3=$b_4 = 0$	2.686173	(16, 79)	0.0020
$H_0^{(3)}$: b_1=b_2=$b_3 = 0$	2.967118	(12, 83)	0.0018
$H_0^{(2)}$: b_1=$b_2 = 0$	3.896184	(8, 87)	0.0006
$H_0^{(1)}$: $b_1 = 0$	6.471353	(4, 91)	0.0001

The $H_0^{(i)}$ test uses the i-th order Taylor expansion ($b_j = 0$ for all $j > i$).

Teräsvirta (1994) Sequential Tests

Null Hypothesis	F-statistic	d.f.	p-value
H_3: $b_3 = 0$	1.080239	(4, 83)	0.3716
H_2: $b_2 = 0 \mid b_3 = 0$	1.249922	(4, 87)	0.2960
H_1: $b_1 = 0 \mid b_2$=$b_3 = 0$	6.471353	(4, 91)	0.0001

All tests are based on the third-order Taylor expansion ($b_4 = 0$).
Linear model is rejected at the 5% level using $H_0^{(3)}$.
Recommended model: first-order logistic.
$\Pr(H_1) \leq \Pr(H_2)$

Escribano-Jorda (1999) Tests

Null Hypothesis	F-statistic	d.f.	p-value
$H_0^{(L)}$: b_2=$b_4 = 0$	1.355888	(7, 79)	0.2358
$H_0^{(E)}$: b_1=$b_3 = 0$	0.893747	(6, 79)	0.5037

All tests are based on the fourth-order Taylor expansion.
Linear model is rejected at the 5% level using H04.
Recommended model: exponential with nonzero threshold.
$\Pr(H_0^{(L)}) < \Pr(H_0^{(E)})$ with $\Pr(H_0^{(L)}) >= .05$

Table 17: Different linearity tests for WQ-input version with smooth transition (STR) model as alternative. The threshold variable is the output gap $y_{t|t}$.

Linearity Tests for STR-BW

Taylor series alternatives: $b_0 + b_1 \cdot s[+ b_2 \cdot s^2 + b_3 \cdot s^3 + b_4 \cdot s^4]$
Threshold variable s: CCPI(-1)

Luukkonen, Saikkonen, and Teräsvirta (1988) Linearity Tests

Null Hypothesis	F-statistic	d.f.	p-value
$H_0^{(4)}$: b_1=b_2=b_3=$b_4 = 0$	1.000100	(16, 79)	0.4656
$H_0^{(3)}$: b_1=b_2=$b_3 = 0$	1.091832	(12, 83)	0.3779
$H_0^{(2)}$: b_1=$b_2 = 0$	0.886902	(8, 87)	0.5310
$H_0^{(1)}$: $b_1 = 0$	1.418112	(4, 91)	0.2343

The $H_0^{(i)}$ test uses the i-th order Taylor expansion ($b_j = 0$ for all $j > i$).

Teräsvirta (1994) Sequential Tests

Null Hypothesis	F-statistic	d.f.	p-value
H_3: $b_3 = 0$	1.463863	(4, 83)	0.2206
H_2: $b_2 = 0 \mid b_3 = 0$	0.393498	(4, 87)	0.8128
H_1: $b_1 = 0 \mid b_2$=$b_3 = 0$	1.418112	(4, 91)	0.2343

All tests are based on the third-order Taylor expansion ($b_4 = 0$).
Linear model is not rejected at the 5% level using $H_0^{(3)}$.

Escribano-Jorda (1999) Tests

Null Hypothesis	F-statistic	d.f.	p-value
$H_0^{(L)}$: b_2=$b_4 = 0$	1.068878	(7, 79)	0.3913
$H_0^{(E)}$: b_1=$b_3 = 0$	0.643613	(6, 79)	0.6950

All tests are based on the fourth-order Taylor expansion.
Linear model is not rejected at the 5% level using $H_0^{(4)}$.

Table 18: Different linearity tests for BW-input version with smooth transition (STR) model as alternative. The threshold variable is inflation $\pi_{t-1|t}$.

Linearity Tests for STR-FW

Taylor series alternatives: $b_0 + b_1 \cdot s[+ b_2 \cdot s^2 + b_3 \cdot s^3 + b_4 \cdot s^4]$
Threshold variable s: Y(+1)

Luukkonen, Saikkonen, and Teräsvirta (1988) Linearity Tests

Null Hypothesis	F-statistic	d.f.	p-value
$H_0^{(4)}$: b_1=b_2=b_3=$b_4 = 0$	3.618718	(16, 79)	0.0001
$H_0^{(3)}$: b_1=b_2=$b_3 = 0$	4.120081	(12, 83)	0.0000
$H_0^{(2)}$: b_1=$b_2 = 0$	5.172401	(8, 87)	0.0000
$H_0^{(1)}$: $b_1 = 0$	9.800630	(4, 91)	0.0000

The $H_0^{(i)}$ test uses the i-th order Taylor expansion ($b_j = 0$ for all $j > i$).

Teräsvirta (1994) Sequential Tests

Null Hypothesis	F-statistic	d.f.	p-value
H_3: $b_3 = 0$	1.688143	(4, 83)	0.1605
H_2: $b_2 = 0 \mid b_3 = 0$	0.681417	(4, 87)	0.6067
H_1: $b_1 = 0 \mid b_2$=$b_3 = 0$	9.800630	(4, 91)	0.0000

All tests are based on the third-order Taylor expansion ($b_4 = 0$).
Linear model is rejected at the 5% level using $H_0^{(3)}$.
Recommended model: first-order logistic.
$\Pr(H_3) \leq \Pr(H_2)$ or $\Pr(H_1) \leq \Pr(H_2)$

Escribano-Jorda (1999) Tests

Null Hypothesis	F-statistic	d.f.	p-value
$H_0^{(L)}$: b_2=$b_4 = 0$	1.432323	(7, 79)	0.2043
$H_0^{(E)}$: b_1=$b_3 = 0$	0.748925	(6, 79)	0.6121

All tests are based on the fourth-order Taylor expansion.
Linear model is rejected at the 5% level using $H_0^{(4)}$.
Recommended model: exponential with nonzero threshold.
$\Pr(H_0^{(L)}) < \Pr(H_0^{(E)})$ with $\Pr(H_0^{(L)}) \geq 0.05$

Table 19: Different linearity tests for FW-input version with smooth transition (STR) model as alternative. The threshold variable is the output gap $y_{t+1|t}$.