NBP Narodowy Bank Polski

NBP Working Paper No. 262

Bagged artificial neural networks in forecasting inflation:

An extensive comparison with current modelling frameworks

Karol Szafranek

NBP Working Paper No. 262

Bagged artificial neural networks in forecasting inflation:

An extensive comparison with current modelling frameworks

Karol Szafranek

Economic Research Department
Warsaw, 2017

Karol Szafranek – Narodowy Bank Polski, Warsaw School of Economics

This research project has been financially supported by the Warsaw School of Economics (decision number KAE/BMN/16/16/16). The views expressed herein are those of their author and do not necessarily reflect those of Narodowy Bank Polski or the Warsaw School of Economics. I would like to thank the anonymous referee for valuable comments. I would also like to thank Paweł Baranowski, Agnieszka Leszczyńska and Grzegorz Szafrański for making the real-time database available for my study as well as Damian Stelmasiak and Grzegorz Szafrański for providing the Polish headline inflation forecasts from the Bayesian vector autoregressive models with Sims-Zha and Villani priors as well as the dynamic factor model included in this paper in the set of benchmarks. Correspondence should be directed to: karol.szafranek@nbp.pl

Published by:
Narodowy Bank Polski
Education & Publishing Department
ul. Świętokrzyska 11/21
00-919 Warszawa, Poland
phone +48 22 185 23 35
www.nbp.pl

ISSN 2084-624X

Contents

Abstract

Accurate inflation forecasts lie at the heart of effective monetary policy. By utilizing a thick modelling approach, this paper investigates the out-of-sample quality of the short-term Polish headline inflation forecasts generated by a combination of thousands of bagged single hidden-layer feed-forward artificial neural networks in the period of systematically falling and persistently low inflation. Results indicate that the forecasts from this model outperform a battery of popular approaches, especially at longer horizons. During the excessive disinflation it has more accurately accounted for the slowly evolving local mean of inflation and remained only mildly biased. Moreover, combining several linear and non-linear approaches with diverse underlying model assumptions delivers further statistically significant gains in the predictive accuracy and statistically outperforms a panel of examined benchmarks at multiple horizons. The robustness analysis shows that resigning from data preprocessing and bootstrap aggregating severely compromises the forecasting ability of the model.

JEL: C22, C38, C45, C53, C55
Keywords: inflation forecasting, artificial neural networks, principal components, bootstrap aggregating, forecast combination

1. Introduction

Forecasting inflation lies at the heart of central bank's activities. Since monetary policy influences the economy via transmission mechanisms with a substantial lag, tackling inflationary or deflationary pressures that may arise in the future requires the policy makers to take decisions ahead of time. Therefore, accurate and reliable inflation forecasts are an indispensable ingredient in formulating monetary policy and targeting future inflation (Orphanides and Wieland, 2008). They play also a vital role in the proper and transparent communication process with the general public (Blinder et al., 2008), since they approximate the shape of the future monetary policy and influence the decision-making process of the private and business sector agents by forming their expectations and affecting negotiations regarding the nominal commitment (Faust and Wright, 2013).

Over the years a significant amount of various approaches has been developed to improve the accuracy of forecasts. In particular, a considerable body of the existing literature is devoted to predicting inflation. The complexity of the price dynamics as well as the plurality of its potential determinants motivate researchers to pursue more and more elastic modelling frameworks that incorporate as much useful information as possible. Employing a large number of potential predictors, methods aimed at reducing the dimensionality of the analysis, shrinkage or forecast combination has systematically become the current state of the art. Several common frameworks should be enumerated here.

Firstly, popular univariate approaches adopted by practitioners include time-series models exploiting the information embedded in large datasets of real activity indicators (Stock and Watson, 1999), asset prices (Forni et al., 2003), survey measures (Ang et al., 2007) or aggregating forecasts of highly disaggregated inflation components (Duarte and Rua, 2007). Secondly, apart from employing large information sets, a strand of researchers develop their models utilizing a generalized Phillips curve framework using Bayesian model averaging across different regression specifications and allowing the coefficients as well as the entire forecasting model to change over time (e.g Groen et al., 2013). Thirdly, multivariate approaches include most often vector autoregressive models augmented with common factors (Berg and Henzel, 2015), shrinkage procedures (Jochmann et al., 2010), time-varying coefficients (Stella and Stock, 2012) or tight steady-state priors (Stelmasiak and Szafrański, 2016a). Fourthly, common structural approaches employ either the univariate present value formulation of the new Keynesian Phillips curve (Rumler and Valderrama, 2010) or the dynamic stochastic equilibrium models (Edge and Gurkaynak, 2010). Finally, an important role play judgemental forecasts (Faust and Wright, 2013), prepared by the professional forecasters and based on applying a subjective filters to the observed developments in the economy. The majority of these approaches have been extensively compared and discussed by Faust and Wright (2013).

With the explosion in the number and the variety of the forecasting methods, attention has also been paid to artificial neural networks – a class of highly non-linear models. Tremendous surge in the research activities took place in the last decade of the twentieth century (Zhang et al., 1998). Although the predictive performance of forecasts generated by the artificial neural networks is ambiguous across the forecast horizons, a number of empirical studies conclude that the improvement in the forecast accuracy upon the standard benchmarks is non-negligible, especially for predicting inflation at longer horizons (e.g. Moshiri and Cameron, 2000; Chen et al., 2001; Nakamura, 2005; Binner et al., 2005) or in times of structural change and considerable uncertainty (McAdam and McNelis, 2005). Moreover, Ahmed et al. (2010) imply that a particular class of these models – single hidden-layer feed-forward artificial neural networks – beside beating linear benchmarks displays also superior predictive performance in comparison to other common machine learning techniques.

Although a myriad of forecasting models and methods are constantly developed, predicting inflation can be still perceived as a daunting challenge. Evidence put forward by Atkeson and Ohanian (2001) suggests that the traditional macroeconomic Phillips curve linking the adjustments in the real activity with the nominal sphere of the economy fails to forecast inflation more accurately than the naive benchmark extrapolating the inflation developments in the most recent past. Stock and Watson (2007) continue this argument underlining that it has become much more difficult to provide a generous value added from a multivariate forecasting model over a univariate benchmark. They show that the U.S. inflation can be well approximated by the univariate unobserved component model where the disturbances follow a stochastic volatility process. The out-of-sample performance of this model is hard to beat. Moreover, in a recent comprehensive comparison of models popularly employed for inflation forecasting Faust and Wright (2013) conclude that judgemental inflation forecasts tend to outperform traditional approaches. The authors argue that this advantage stems from the fact that a professional forecaster can elastically choose two boundary values denoting the current and the expected future local mean of inflation and simply employ a smooth transition path to connect these two extreme states.

This difficult challenge of inflation forecasting continues. Recently headline inflation in many countries has revealed a peculiar behaviour throwing down the gauntlet to both models and professional forecasters. After 2011 the inflation rate across most of the highly developed economies and a number of emerging ones has abruptly decreased and remained persistently low for a prolonged period of time, deviating significantly from the official targets of the central banks. A protracted period of low inflation has been observed also in Poland (Szafranek, 2017). As as result, most of the stationary models displaying a mean-reverting property has failed in forecasting inflation for longer horizons and revealed a significant bias.

This puzzling development of inflation motivates me to study whether employing a large combination of highly non-linear models can improve inflation forecasts in Poland upon the battery of benchmarks. For this purpose, I employ the Polish data. In the forecasting exercise I consider several novelties by studying the joint effect of employing factor analysis, bootstrap aggregating and forecast combinations in a thick-based approach of inflation forecasting. In a pseudo real-time out-of-sample experiment I study the usefulness of combining a large number of forecasts generated from bagged single hidden-layer artificial neural networks which architecture is governed by the realization of a random variable. More specifically, in the first step I employ a vintage, large macroeconomic database and extract several unobserved factors using principal components analysis. I treat these derived components as input variables for each of the 10 000 individual single hidden-layer feed-forward artificial neural networks. Secondly, in order to avoid subjectivity in choosing the number of units in the hidden layer, I assume that it follows a zero-truncated Poisson distribution. Thirdly, I apply the supervised learning procedure to train the models and employ bootstrap averaging to stabilize the results of the whole model and improve its forecast accuracy. Next, I use combination methods to aggregate all individual predictions into the point combination forecast. Lastly, I perform an extensive comparison between the competing models and report the outcomes of the sensitivity analysis regarding the key parameters of the model.

The following conclusions are warranted in the paper. First, the combined forecast from a large number of bagged artificial neural networks is able to outperform a battery of competing models for longer horizons. Importantly, this superior predictive ability is more pronounced once the inflation deviates significantly from its official target and remains persistently low. This divergence in the forecast accuracy of the considered models and the proposed approach stems from the fact that most time-series models and judgement forecasts available for the Polish headline inflation reflect a strong property of mean-reversion. As a result, they have failed to pin down the unorthodox behaviour of the headline inflation after 2011. Secondly, according to the Giacomini and White (2006) test statistics, the gains in the forecast accuracy are statistically significant in a number of cases. Thirdly, combining best performing linear and non-linear approaches delivers further gains in the forecast accuracy and statistically outperforms a panel of univariate and multivariate approaches, including the Atkeson and Ohanian (2001) benchmark. Lastly, the robustness analysis indicates that a parsimonious architecture of the neural network is preferred and applying proper data preprocessing and the bootstrap aggregating greatly increases the forecast accuracy.

The outline of the paper is as follows. Section 2 briefly introduces the concept of an artificial neural network model and reviews the recent findings from the literature on the predictive ability of this class of non-linear models. The model, the data and the estimation process is described in detail in section 3. Main results are presented in section 4 along with an extensive robustness check. Section 5 concludes.

2. Literature review

Artificial neural networks are one of the most powerful state-of-the-art machine learning frameworks (Ahmed et al., 2010). In this paper I restrict my attention to the simplest representative of this class of models: a single hidden-layer feed-forward artificial neural network. It can be mathematically represented as follows:

$$\hat{y_t} = f\left(\sum_{j=1}^{J} \hat{\nu}_j g(\hat{\omega}_j^{'} x_t)\right) \tag{1}$$

where $x_t^{'} = (1, x_{1,t}, \ldots, x_{n,t})^{'}$ are explanatory (input) variables at time t, $\hat{\omega}_j^T = (\hat{\omega}_{1,j}, \ldots, \hat{\omega}_{n+1,j})$ is the vector of estimated weights connecting the input variables with the jth hidden node, $j = (1, \ldots, J)$, $\hat{\nu}_1, \ldots, \hat{\nu}_J$ are the estimated weights for the output node and $\hat{y}_t$ is the network output at time t. The functions g and f represent the bounded activation function for the neurons in the hidden and the output layer, respectively. Naturally, more complex structures exist (e.g Schmidhuber, 2015; Prieto et al., 2016), but they remain beyond the scope of interest of this paper.

Using the neural network terminology, the architecture of the model is constituted by the number of layers and neurons (nodes, units) grouped in each layer. A single hidden-layer feed-forward artificial neural network consists of three layers: the input layer (with the number of neurons corresponding to the number of explanatory variables), the hidden layer (with the neurons transforming the incoming information most often via a continuous, monotonic and differentiable activation function) and the output layer (with the number of neurons corresponding to the number of dependent variables). The flexibility of the model and the ability to reproduce non-linear patterns is determined by the hidden layer. Hornik et al. (1989) show that under mild conditions one hidden layer with a sufficient number of neurons in the hidden layer provides an approximation to any piecewise continuous function to any desired level of precision.

In a single hidden-layer feed-forward artificial neural network all neurons are fully connected between layers by directed arcs. As a result, information runs exclusively from the input to the output layer (hence the network is feed-forward). The weights determine the relative importance of the flowing information by attenuating or amplifying the flowing signal. They are estimated in an iterative process with the use of the training algorithms, so that the network is said to learn the relationship between the presented patterns. During the learning process the information is propagated through the network with initially randomized weights. In the output layer the predicted outcome is compared with the target value and the parameters are adjusted in the direction that minimizes the chosen cost function. Next, the modified weights are used to recalculate the predicted value and compare it again with the target value. The learning process continues until the network reaches a specified criterion.

The tremendous surge in the research activities studying the applications of the artificial neural networks for forecasting financial and macroeconomic variables has been observed in the last decade of the twentieth century. From the forecasting perspective, Zhang et al. (1998) provide an extensive summary on the theoretical considerations regarding the construction of the neural networks, highlight key uncertainties connected with the modelling process as well as summarize numerous empirical applications in forecasting. The authors stress that some unique characteristics of the artificial neural networks – non-linearity, flexibility and adaptability – make them suitable for forecasting purposes. However, they underline that their predictive performance remains inconclusive and may be problem-specific. They state that in order to properly utilize this class of models, some key limitations have to be acknowledged. These shortcomings include a large degree of uncertainty in the modelling phase since choosing proper network architecture and learning parameters is often based on a trial-and-error methodology, the poor interpretability of the results due to their highly non-linear nature (though Refenes et al. (1994) propose a sensitivity analysis for the weights aimed at understanding the relative importance of the network's inputs) and issues of overfitting and local convergence (Goffe et al., 1994) connected with the necessity to estimate a large number of parameters. They conclude, however, that artificial neural networks are a promising alternative to traditional linear methods. A more general perspective on the artificial neural networks, including also recent developments in this field, provide Prieto et al. (2016).

The overall verdict regarding the accuracy of the inflation forecasts generated by the artificial neural networks remains ambiguous. However, several empirical applications confirm that these models can generate competitive forecasts in comparison to the traditional approaches, especially for longer horizons or in periods of significant structural change and uncertainty.

Several applications point to the superior predictive accuracy of these models. Swanson and White (1997) conclude that the artificial neural networks appear to be a useful alternative to less flexible, fixed specifications linear models, particularly at forecast horizons greater than one-step-ahead. Moshiri et al. (1999) compare forecasts from three artificial neural network models (feed-forward, recurrent and radial-basis neural network) with the econometric time-series approaches and conclude that the non-linear models outperform the linear benchmarks especially at the 12-month horizon. Among the non-linear models, a simple feed-forward artificial neural network tends to perform the best. Moshiri and Cameron (2000) provide further evidence claiming that a simple backpropagation neural network delivers as accurate forecasts as its traditional competitors – a structural, a vector autoregressive and Bayesian vector autoregressive model – and in some cases they are able to outperform them. Again, the performance of the non-linear model is best for the twelve-month-ahead horizon. Chen et al. (2001) examine the forecasting performance of the stylized Phillips curve models and the semiparametric autoregressive neural network with exogenous variables for the U.S. inflation. They find that the latter models outperform linear models in one-step-ahead forecast competition. Nakamura (2005) concludes that the artificial neural network outperforms the univariate autoregressive models for forecast horizons not larger than 3 quarter. The author highlights that not accounting for the risk of overfitting of the artificial neural network by implementing early stopping mechanism can severely compromise the predictive accuracy of the model. McAdam and McNelis (2005) show that a neural network-based thick model based on the Phillips curve formulation outperforms the linear models for the real-time and bootstrap forecasts. They claim that the pay-off of utilizing the artificial neural networks in the thick modelling framework is visible in the period of structural change and uncertainty. Binner et al. (2005) investigate the relative performance of the Euro area inflation forecasts generated by the artificial neural networks and the linear univariate and multivariate models. They conclude that both the within-sample and the out-of-sample performance of non-linear models is superior.

Contrary evidence is also present in the empirical literature. Teräsvirta et al. (2005) show overall mixed results for the forecasting performance of inflation and other macroeconomic variables in seven economies by the single hidden-layer artificial neural networks. However, they conclude that at long horizons the forecast accuracy improves. Moreover, they claim that combining non-linear models (smooth threshold autoregressive models with artificial neural networks) with uneven performance leads to remarkably good inflation forecasts. Binner et al. (2006) provide evidence that non-linearities in the U.S. inflation rate in the short-term are accounted equally well by a Markov-switching model and a combination of recurrent neural networks, but in the long-term the Markov-switching approach is preferred. Kock and Teräsvirta (2013) find that artificial neural networks do not outperform a linear autoregressive approach at any horizon for the Finnish inflation rate. However, they utilize only automated techniques.

Not limiting the review of the artificial neural network applications to inflation forecasts, Ahmed et al. (2010) show in a large-scale comparison based on the M3 time-series competition data that the most popular artificial neural networks – single hidden-layer feed-forward models – outperform other common machine learning methods. They highlight that different preprocessing methods significantly impact the performance of the model. Furthermore, Andrawis et al. (2011) conclude that a simple average forecast combination of the best linear and non-linear approaches delivers superior results.

Taking into account the advantages as well as the caveats of the artificial neural networks, in this paper I combine several methods of improving the predictive ability put forward in the literature with the artificial neural networks and examine their joint accuracy in forecasting the Polish inflation rate in a period of systematically falling and persistently low inflation.

3. The model and the data

This section provides the description of my forecasting exercise. Firstly, I present the specification of the non-linear model employed throughout the paper: a single hidden-layer feed-forward artificial neural network. Secondly, I describe the vintage dataset and the preprocessing step. Next, I discuss the learning procedure which comprises of employing bootstrap aggregating and a specific learning algorithm. Then I establish several simple methods used to obtain the combination forecast, introduce shortly a round-up of competing models utilized for comparison purposes and discuss the forecast evaluation criteria. I end this section by presenting a number of alternative specifications of the model considered as a robustness check.

3.1. The model

As outlined in the previous section, a considerable strand of literature suggests that a single hidden-layer artificial neural network provides the most accurate forecasts. Therefore, in this paper I focus my attention on this class of non-linear models. Equation (2) presents its mathematical representation:

$$y_t = h\Bigg(\sum_{j=1}^{q} \beta_j g\big(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} x_{i,t}\big)\Bigg) + \epsilon_t \tag{2}$$

where: the output of the artificial neural network y_t at time t depends on the set of explanatory variables $x_t = (x_{1,t}, \ldots, x_{p,t})$, the bias (intercept) of the jth hidden layer neurons β_{0j}, the model connection weights β_j, $j = (1, \ldots, q)$ and β_{ij}, $i = (1, \ldots, p)$, $j = (1, \ldots, q)$ corresponding to the arcs (synapses) between neurons (units). In this framework p is the number of the input nodes equal to the number of the explanatory variables, q denotes the number of the hidden units, g is the bounded activation function in each neuron of the hidden layer and h is the activation function of the output layer. The error term ϵ_t is assumed to be $IID(0, \sigma^2)$. In this particular exercise y denotes the monthly headline inflation rate in Poland.

Building an artificial neural network requires determining its architecture. At the beginning, explanatory (input) variables have to be selected. I discuss this choice in section 3.2. Secondly, one has to choose the number of neurons (nodes) in the hidden layer and their activation (squashing) function. Although some algorithms have been proposed in the literature (e.g. Zhang et al., 1998; Moshiri and Cameron, 2000; Teräsvirta et al., 2005), the choice of the number of neurons in the hidden layer remains problematic and normally is a subject to a rule-of-thumb methodology and tedious, time-consuming experiments. Most authors examine different specifications and select the one yielding the best results, which may introduce subjectivity into the discussion of the outcomes. In this paper I present a thick modelling approach by estimating a large number of artificial neural networks. This framework allows me to remain agnostic about the number of the neurons in the hidden layer. Instead, I account for the uncertainty regarding the optimal number of the neurons in the hidden layer by assuming that it is a random variable following a zero-truncated Poisson (Ztp) distribution:

$$ZTP(J) = P(Q = q \,|\, Q > 0) = \frac{\lambda^q}{(e^{\lambda} - 1)q!} \tag{3}$$

with the expected value equal to:

$$\gamma = E[Q] = \frac{\lambda}{1 - e^{-\lambda}} \tag{4}$$

where λ is the parameter of the standard Poisson distribution. This choice is motivated twofold. First, I want to restrict all artificial neural networks entering the thick model to be non-linear. Hence, each model must possess at least one unit in the hidden layer. Second, as the number of the neurons in the hidden layer increases, the necessity to estimate a large number of weights (parameters) arises. This may lead to the loss of generalization properties and overfitting problems. Hence I favour models that are rather parsimonious and choose a rightly skewed distribution. In the baseline specification of the model I choose

$\gamma = 3$. In the paper I examine how different values of the parameter γ governing the expected size of the artificial neural network influence the forecasting ability of the model.

The information in each neuron of the hidden layer is transformed by an activation function g, which ideally should be continuous, non-decreasing and differentiable. Popularly, a logistic function is chosen. However, since inflation may take negative values, I follow the approach of Binner et al. (2005) as well as Nakamura (2005) and utilize the hyperbolic tangent to process the linear combination of explanatory variables entering each hidden layer neuron. Moreover, Bishop (1995) states that this function gives rise to faster convergence of the learning algorithms. Therefore, the linear combination $\hat{u}_{j,t} = \hat{\beta}_{0j} + \sum_{i=1}^{p} \hat{\beta}_{ij} x_{i,t}$ of input variables and weights entering the jth unit of the hidden layer is transformed according to the following equation:

$$tanh(\hat{u}_{j,t}) = \frac{e^{2\hat{u}_{j,t}} - 1}{e^{2\hat{u}_{j,t}} + 1} \tag{5}$$

For the output unit I consider the identity function: $h(\sum_{j=1}^{q} \hat{\beta}_j g(\hat{u}_{j,t})) = \sum_{j=1}^{q} \hat{\beta}_j g(\hat{u}_{j,t})$.

3.2. The data and the preprocessing stage

My specific forecasting experiment concerns the Polish headline inflation. For that purpose I utilize the real-time dataset of 188 potential predictors spanning the period from January 1999 to December 2016. During the exercise I generate a total number of 72 quasi real-time forecasts in a recursive manner starting in January 2011. I asses inflation forecast accuracy over the maximal horizon of 12 months.

The vintage database consists of monthly time series used in the study of forecasting inflation with the dynamic factor model by Baranowski et al. (2010). For clarification purposes, let me here denote the data available in each vintage by $\mathbb{X}_v = \{x_{it}\}$, $i = 1, \ldots, N$, $t = 1, \cdots, T_v$ and $v = 1, \ldots, 72$, where $N = 188$ stands for the number of all potential explanatory variables and T_v denotes the size of the vth vintage. I apply a recursive strategy using an expanding estimation window. Therefore, the first (last) vintage of data spans the period 1999:01-2010:12 (1999:01-2016:12) and contains 144 (216) observations. The utilized variables depict a broad illustration of the domestic and foreign stance of the economy and can be categorized into the following groups: market conditions in trade, industrial production and construction (48), foreign trade (10), exchange rates (4), money market and bond yields (7), stock exchange indices (8), money supply (6), official reserve assets (3), consolidated balances of the monetary institutions (11), PPI indices (9), labour market indices (10), public finance indices (4), international commodity prices (15), fuel prices (3), domestic agriculture products prices (6), CPI indices (36) and miscellaneous (8). A detailed list of all variables used in the analysis is available on request.

Preprocessing of the data is often found to be crucial for the forecasting performance of the artificial neural networks and is commonly adopted in the literature (e.g. Zhang et al., 1998; McAdam and McNelis, 2005; Ahmed et al., 2010). Preprocessed data decrease the risk of computational problems and make the training process more efficient (Zhang et al., 1998). Moreover, utilizing all potential predictors is improbable due to a large number of weights requiring estimation.

Taking this into consideration, I assume that the information in each vintage can be effectively compressed to several ($r \ll N$) common factors. Following the frequent practice (e.g. Stock and Watson, 1999) in the first step I seasonally adjust all variables in the quasi real-time. This is done to avoid extracting a common for explanatory variables seasonal pattern which can significantly depart from the seasonal pattern of the headline inflation (Baranowski et al., 2010). Moreover, though artificial neural networks should be able to learn non-linear patterns, Nelson et al. (1999) state that using seasonally adjusted data improves the overall forecast accuracy of the neural networks.

In the next step, I vertically align the data due to their different publication lags which results in the so-call ragged-edge problem. Thirdly, I derive a small number ($r \ll N$) of the largest factors $\hat{f}_t = (\hat{f}_{1,t}, \ldots, \hat{f}_{r,t})$ using the principal components analysis on the standardised data and treat them as the explanatory (input) variables. Moreover, since the principal components are estimated on the seasonally-adjusted data and the dependent variable reveals a clear seasonal pattern, I extract the seasonal component s_t from the monthly headline inflation using a X-13-ARIMA-SEATS filter and add the estimates of the seasonal pattern to the set of explanatory variables in order to account for the seasonality in the dependent variable. Finally, following McAdam and McNelis (2005) as well as Ahmed et al. (2010) I scale all variables (the dependent variable, the common factors and the seasonal pattern), adjusting them to the range $[0 \div 1]$ according to the following formula:

$$\upsilon^*_{k,t} = \frac{\upsilon_{k,t} - min(\upsilon_k)}{max(\upsilon_k) - min(\upsilon_k)} \tag{6}$$

Here υ denotes the variable of interest, $\upsilon \in \{y, f, s\}$.

3.3. Estimation

In this section I discuss in detail the estimation methodology. Firstly, I outline the bootstrap aggregating procedure used in my framework. Secondly, I describe the learning algorithm.

3.3.1. Bootstrap aggregating

Considerable uncertainty regarding the proper architecture of the artificial neural network as well as the serious risk of overfitting non-linear models motivates me to use the thick modelling approach (Granger and Jeon, 2004). In simple terms, I utilize a substantial number of alternative specifications instead of focusing my attention on a dominant model and discarding other potential specifications. Thick modelling in the context of artificial neural networks has been successfully adopted by McAdam and McNelis (2005) who combine forecasts of twenty neural networks with different architectures. They conclude that pooling the forecast even from a rather small set of models can have superior predictive accuracy. This approach can be easily extended to account for a wider variety of specifications which I study in this paper.

I introduce thick modelling in my framework by employing bootstrap aggregation (Breiman, 1996). In short, bagging consists of perturbing the learning sample, estimating the model on several pseudo-new bootstrap samples, peforming the forecasts and averaging over all obtained outcomes. Breiman (1996) shows that bagging reduces the variance of the full model and greatly improves the accuracy of methods for which the outcomes heavily depend on the particular observed sample. Since artificial neural networks are data-driven and can suffer high variance in the estimation (Zhang et al., 1998), bagging may substantially increase their predictive performance and reduce the forecast dispersion.

Several applications of bagging reveal that this approach can significantly decrease the mean squared-errors of predictions. Inoue and Kilian (2008) show that regression-based bagged models can deliver substantial increase (up to around 40%) in the forecasting accuracy of the U.S. inflation rate relative to the autoregressive benchmark, especially at a 12-month horizon. However, the authors conclude that other methods – the Bayesian shrinkage, the ridge regression, the iterated LASSO and the Bayesian model average predictor – are capable of achieving similar gains in the forecast accuracy. Rapach and Strauss (2010) claim that bagging forecasts are often more accurate than a variety of combination forecasts. They suggest that a conjunction of bagging and combination forecasts can deliver further improvements in the predictive accuracy. Khwaja et al. (2015) show that forecasts from bagged artificial neural networks are characterized by lower forecast errors and variance in the forecast values than a variety of considered models.

Before applying bagging, in order to achieve generalization properties, i.e a satisfactory performance of the model on unseen data, in machine learning approaches the dataset is normally split into three parts: the learning, the validation and the test sample. The neural network recognizes the relationship between the input patterns and the target variable on the training sample. Its performance is evaluated on the validation set disjoint from the training set. However, since early stopping is implemented to avoid overfitting, the validation sample takes part indirectly in the learning process, as the training stops once the minimum error is reached on the validation set and not learning set. Therefore, to obtain an unbiased estimate of the generalization capacity of the artificial neural network forecasts on the test set are necessary.

To apply bootstrap aggregating, I start by arranging the data in the vth vintage in a set of tuples $\{y_t, f_t^{'}, s_t\}$. They form a matrix of dimension $T_v \times r + 2$:

$$\begin{matrix} y_1 & \hat{f}_1^{'} & s_1 \\ \vdots & \vdots & \vdots \\ y_{T_v} & \hat{f}_{T_v}^{'} & s_{T_v} \end{matrix}$$

where: y denotes the dependent variable (seasonally unadjusted monthly headline inflation), $\hat{f}$ denotes the $r \times 1$ factor estimates, s denotes the real-time estimate of the seasonal component of inflation derived using an X-13-Arima-Seats, $t = 1, \ldots, T_v$ denotes time and T_v denotes the number of observation for a vth vintage. Next, I follow a common practice in the machine learning approaches (Zhang et al., 1998) and for a given vintage v I construct a learning set $\mathbb{L}_v$ and a validation set $\mathbb{V}_v$ by randomly distributing the tuples. In the baseline specification of the model, the learning set contains $\delta_1 = 70\%$ of all tuples of a given vintage (the dimension of $\mathbb{L}_v$ is $L_s \times r + 2$, $L_s = \delta_1 \times T_v$ and the validation set contains the remaining $1 - \delta_1 = \delta_2 = 30\%$ of all tuples (the dimension of $\mathbb{V}_v$ is $V_s \times r + 2$, $V_s = \delta_2 \times T_v$. The test set includes the forecasting period and its dimension is $H \times r + 2$, $H = 12$.

The training set $\mathbb{L}_v$ for the vth vintage is then bagged. Breiman (1996) shows that the gain from the use of bagging stabilizes if the number of bootstrap pseudo-samples B exceeds 25, but Inoue and Kilian (2008) as well as Rapach and Strauss (2010) use 100 replicates instead. I follow their approach and generate $B = 100$ bootstrap pseudo-new learning sets of size L_s equal to the initial learning set $\mathbb{L}_v$ by sampling uniformly with replacement. This guarantees the independence between samples with the probability of selecting a particular tuple to the training set amounting to around 63% (Breiman, 1996). Each bagged training set $\mathbb{L}_{v,i}$, $i = 1, \cdots, B$ for a given vintage v is a matrix of dimension $L_s \times r + 2$ and has the following structure:

$$\begin{matrix} y_1^* & \hat{f}_1^{'*} & s_1^* \\ \vdots & \vdots & \vdots \\ y_{L_s}^* & \hat{f}_{L_s}^{'*} & s_{L_s}^* \end{matrix}$$

In the baseline specification of the model, for each of the pseudo-new bootstrapped learning sets $\mathbb{L}_{v,i}$ I estimate $D = 100$ models defined by equation (2) with varying architectures determined by the realization of the random variable Q. As a results, in total I obtain $B \times D = 10\,000$ estimated artificial neural networks.

The validation set is not bagged in order to assess the out-of-sample quality on the historical data of all neural networks estimated on all generated training samples. Therefore, it has the following structure:

$$\begin{matrix} y_1^{**} & \hat{f}_1^{'**} & s_1^{**} \\ \vdots & \vdots & \vdots \\ y_{V_s}^{**} & \hat{f}_{V_s}^{'**} & s_{V_s}^{**} \end{matrix}$$

I evaluate the quality for the jth artificial neural network on the validation set using a simple mean squared error statistic: $MSE_j = \sum_{i=1}^{V_s}(y_i^{**} - \hat{y}_{i,j}^{**})$, where $\hat{y}_{i,j}^{**}$ denotes the prediction implied by the estimated jth model using the historical data in the validation set.

The test set consists of the predicted factors and the predicted seasonal component. To obtain estimates for the test set I assume that the r common factors f_t are generated by the following multivariate autoregressive process in the spirit of Bernanke et al. (2005):

$$\hat{f}_t = \mu_0 + \sum_{j=1}^{p_f} \mu_j \hat{f}_{t-j} + \nu_t \tag{7}$$

where: the maximal lag p_f is determined using the Bic criterion. The model is estimated on the stationary data before any data splitting. Next, I generate forecasts of the common factors in an iterative manner:

$$\hat{f}_{t+h} = \hat{\mu}_0 + \sum_{j=1}^{p_f} \hat{\mu}_j \hat{f}_{t+h-1} \tag{8}$$

The seasonal component is estimated using the X-13-Arima-Seats decomposition of the predicted monthly headline inflation in the forecast period. As the forecast is an out-of-sample experiment, the test set does not contain the historical realization of inflation. Instead, only the theoretical values implied by the Favar model and the seasonal model enter the test set:

$$\begin{matrix} \hat{f}_1'^{P} & s_1^P \\ \vdots & \vdots \\ \hat{f}_{12}'^{P} & s_{12}^P \end{matrix}$$

Each of the trained artificial neural network is utilized to generate a forecast of inflation in a twelve-month horizon y_{T_v+h}, $h = 1, \ldots, 12$. The out-of-sample forecast of each of the 10 000 artificial neural network is performed by treating the theoretical values from the test set as explanatory variables in the forecast period, feeding them to the model (2) and calculating the output using the estimated weights.

3.3.2. Learning procedure

In order to determine the weights, a training algorithm needs to be employed. For each artificial neural network I use the resilient propagation algorithm proposed by Riedmiller and Braun (1993). It is a robust, fast and stable adaptive gradient-based optimization technique used for the supervised learning (Igel and Hüsken, 2003). It eliminates the harmful influence of the size of the partial derivative on the weight step and considers instead only the sign of the derivative as an indication of the direction of the weight update. Moreover, it allows me to evade the arbitrary choice of the learning and momentum rates, which are present in other standard learning algorithms.

Supervised learning algorithm relies on the numerical optimization of the artificial neural network's weights in order to minimize an error function, which commonly takes into account the sum of squared differences between the theoretical and empirical values. In simple terms, the input pattern (the explanatory variables) is presented to the network, propagated forward until it reaches the output layer where the outcome of the network (the theoretical or the predicted values) is compared with the teaching input (the empirical values). The error of the model is then either accepted or rejected, in which case the connection weights are updated and the procedure is repeated.

In order to mitigate the risk of overfitting, in the paper I establish a composite error function that accounts for an additional ingredient – a regularisation component that is aimed at balancing between the recognition and generalization ability of the artificial neural network. It penalizes the model with exceptionally large weights which results in better generalization properties. I use a fairly standard error function defined as a sum of the squared residuals and the penalty term:

$$E = \sum_{i=1}^{L_s} (\hat{y}_i^* - y_i^*)^2 + 10^{-\theta} \Big(\sum_{j=1}^{q} \sum_{i=0}^{p} \beta_{ij}^2 + \sum_{j=1}^{q} \beta_j^2 \Big) \tag{9}$$

where $l = 1, \ldots, L_s$ indexes the number of observations in the considered training sample and $\hat{y}^*$, y^*, θ, β_{ij}, β_j denote the predicted values of the model, the empirical values, the regularization decay parameter, the weights between the ith input and the jth hidden neuron of the network and the weights between the jth hidden neuron and the output, respectively.

For simplification purposes, let me assume for now that β_{ij} denotes the weight between the ith and the jth neuron (unit), irrespective of the layer. For each single hidden-layer feed-forward artificial neural network, I initialize the learning process by randomizing the weights β_{ij} of the network assigned to all arcs between all neurons in the range $[-0.5 \div 0.5]$. During the learning process each weight β_{ij} determining the strength of the signal flowing from the ith to the jth neuron is iteratively modified according to the following equation:

$$\beta_{ij}^{(k+1)} = \beta_{ij}^{(k)} + \Delta\beta_{ij}^{(k)} \tag{10}$$

where k and $\Delta\beta_{ij}$ denote the learning iteration and the individual update for weight β_{ij}, respectively.

In the resilient propagation, the sign of the update depends on the sign of the partial with respect to the weight derivative of the error measure E. For increasing (decreasing) errors the weights is decreased (increased) with the update value. However, if the partial derivative changes sign, the previous weight-update is reverted. Hence, the update value takes the form:

$$\Delta\beta_{ij}^{(k)} = \begin{cases} -\Delta_{ij}^{(k)} & \text{if} \quad \frac{\partial E^{(k)}}{\partial \beta_{ij}} > 0 \\ +\Delta_{ij}^{(k)} & \text{if} \quad \frac{\partial E^{(k)}}{\partial \beta_{ij}} < 0 \\ 0 & \text{else} \end{cases} \tag{11}$$

The individual step-size Δ_{ij} is calculated as follows:

$$\Delta_{ij}^{(k)} = \begin{cases} \min(\eta + * \Delta_{ij}^{(k-1)}, \Delta_{max}) & \text{if} \quad \frac{\partial E^{(k-1)}}{\partial \beta_{ij}} * \frac{\partial E^{(k)}}{\partial \beta_{ij}} > 0 \\ \max(\eta^{-} * \Delta_{ij}^{(k-1)}, \Delta_{min}) & \text{if} \quad \frac{\partial E^{(k-1)}}{\partial \beta_{ij}} * \frac{\partial E^{(k)}}{\partial \beta_{ij}} < 0 \\ \Delta_{ij}^{(k-1)} & \text{else} \end{cases} \tag{12}$$

where $0 < \eta^{-} < 1 < \eta^{+}$. The adaptation rule takes into account whether the partial derivative with respect to the weight β_{ij} changes its sign in the consecutive steps. If so, the step-size is decreased by the factor η^{-} as the previous update could have been too large resulting in missing the local minimum. If not, the step size is increased by factor η^{+} to accelerate the speed of convergence. The adjustment of the step-size Δ_{ij} is bounded by the parameters Δ_{min} and Δ_{max}.

The learning process of the artificial neural network stops and the model becomes trained once a certain termination criterion is fulfilled (e.g. the number of the iterations exceeds a predefined value, the early-stopping is implemented or the relative error tolerance on the training or the validation set is met).

In the baseline specification of the model I specify the starting values of the parameters in accordance with the Riedmiller and Braun (1993) approach. Hence, the individual step-size is $\Delta_{ij} = 0.1$, the boundary parameters equal $\Delta_{min} = 1e^{-6}$ and $\Delta_{max} = 50$, η^{+} and η^{-} are fixed for the learning algorithm to 1.2 and 0.5, respectively. The authors claim that the choice of this parameters is not critical and does not influence the convergence time. As the error measure includes a regularisation component, the decay parameter θ is set to 2. In the robustness check I also consider alternative values of the parameters Δ_{ij} and θ.

3.4. Forecast combinations

To further improve the forecasting accuracy I combine the point forecast from all individual models using several simple methods. A significant strand of literature points to the superiority of the forecast combinations over the predictions from the thin modelling framework (Timmermann, 2006). Moreover, numerous empirical studies exemplify the phenomenon of the forecast combination puzzle stating that simple combination schemes disregarding the correlation between forecast errors are often superior in comparison to refined methods estimating the theoretically optimal combination weights. To name an example, in a comprehensive study Stock and Watson (2004) conclude that forecasts generated with the use of simple combination schemes, i.e. arithmetic mean, median, weights based on inverse mean squared forecast error or weights experiencing a very little time variation provide largest gains in accuracy. Broadly speaking, equal weights set a surprisingly difficult benchmark to beat (Timmermann, 2006), but the reason for it is not entirely understood at the current juncture.

I perform forecast averaging on the test set. Following the key results from the empirical literature, in my paper I do not compare elaborate forecast combination schemes. Instead I focus on five simple methods of pooling individual forecasts of all single hidden-layer feed-forward artificial neural networks for a given vintage. These include the mean forecast $\hat{y}^M$, median forecast $\hat{y}^{MD}$, trimmed mean forecast $\hat{y}^{TM}$ and forecast with weights depending inversely on the performance of each individual model on the learning set $\hat{y}^L$ and the validation set $\hat{y}^V$. Hence, the combination schemes are given by:

$$\hat{y}^M_{T_v+h} = \frac{1}{n}\sum_{j=1}^{n}\hat{y}^j_{T_v+h} \tag{13}$$

$$\hat{y}^{MD}_{T_v+h} = \text{median}\{\hat{y}^{(1)}_{T_v+h}, \ldots, \hat{y}^{(n)}_{T_v+h}\} \tag{14}$$

$$\hat{y}^{TM}_{T_v+h} = \frac{1}{n-2k}(\hat{y}_{T_v+h,k+1} + \hat{y}_{T_v+h,k+2} \cdots \hat{y}_{T_v+h,n-k}) \tag{15}$$

$$\hat{y}^L_{T_v+h} = \sum_{j=1}^{n} w^L_j \hat{y}^j_{T_v+h} \tag{16}$$

$$\hat{y}^V_{T_v+h} = \sum_{j=1}^{n} w^V_j \hat{y}^j_{T_v+h} \tag{17}$$

where $\hat{y}^j$ is the forecast from the individual, jth model, $n = B \times D$, h is the forecast horizon, k denotes the 5% symmetric trimming following Stock and Watson (2003) and the weights w^L_j and w^V_j based on the relative performance on the training and the validation set and are calculated as follows:

$$w^L_j = \frac{m^{-1}_{j,L}}{\sum_{j=1}^{n} m^{-1}_{j,L}} \tag{18}$$

$$w^V_j = \frac{m^{-1}_{j,V}}{\sum_{j=1}^{n} m^{-1}_{j,V}} \tag{19}$$

$m_{j,L} = \sum_{i=1}^{L_s}(y^*_i - \hat{y}^*_{i,j})$ and $m_{j,V} = \sum_{i=1}^{V_s}(y^{**}_i - \hat{y}^{**}_{i,j})$. The combination forecasts are then compared with traditional linear approaches.

3.5. A round-up of forecasting models

To establish the performance of the proposed model, I compare its prediction accuracy across horizons with a battery of the short-term Polish headline inflation forecast from the univariate and multivariate, linear and non-linear models. As forecast combinations can have superior predictive accuracy compared to their constituent parts I also combine the best performing alternative approaches with equal weights to obtain a combined forecast and examine, if pooling the predictions from several different approaches delivers further gains in the forecast accuracy. The monthly and the yearly headline inflation is denoted as y_t and y^{yy}_t, respectively. The following list summarizes the models competing in the forecasting horse race:

1. The pure random walk model for the seasonally adjusted monthly headline inflation (RW). The forecast for horizon h is obtained using the following rule: $y_{T_v+h} = y_{T_v}$.
2. The random walk model for the monthly headline inflation closely related to the Atkeson and Ohanian (2001) specification (AO). The forecast for horizon h is generated by the following rule: $y_{T_v+h} = \frac{1}{12}\sum_{j=1}^{12} y_{T_v-j+1}$.
3. The recursive autoregressive process of order one (AR1). The model has the following form: $y_t = \rho_0 + \rho_1 y_{t-1} + \epsilon_t$ and is estimated using the whole information for the seasonally adjusted headline inflation in the vth vintage. The forecast is constructed by iterating in the recursive manner the one-step-ahead forecasts.
4. The autoregressive process of order twelve (AR12). The model has the following form: $y_t = \rho_0 + \sum_{i=1}^{12}\rho_i y_{t-i} + \epsilon_t$ and is estimated using the whole information for the headline inflation in the vth vintage. Similarly to the AR1 model, the forecast is constructed using an iterative process to obtain the prediction for h periods ahead.

5. The seasonal autoregressive moving average model (BS). The specification of the model is as follows: $\phi(L)\Phi(L^s)(1-L)^d(1-L^s)^D y_t = c + \theta(L)\Theta(L^s)\epsilon_t$, where: y_t is the monthly headline inflation, c is the mean term, ϵ_t is the independent disturbance, L is the backshift operator, s is the seasonal cycle, $\phi(L)$ and $\Phi(L^s)$ are the lag polynomials for the autoregressive part and the seasonal autoregressive part, $\theta(L)$ and $\Theta(L^s)$ are the lag polynomials for the moving average and the seasonal moving average, $(1-L)^d$ and $(1-L^s)^D$ are the differencing operators. The exact specification of the $ARIMA(p,d,q)(P,D,Q)_s$ model in each vintage is optimized using the Bic. The model is estimated on all available information for the headline inflation in the vth vintage. The forecast is a dynamic out-of-sample.
6. The factor augmented autoregression (FAV). The model has the following representation: $\xi_t = \nu_0 + \sum_{j=1}^{p_f} \nu_j \xi_{t-j} + \varsigma s_t + \epsilon_t$, where $\xi_t = (y_t, \hat{f}_{1,t}, \ldots, \hat{f}_{r,t})'$, $\{\hat{f}_{i,t}\}_{i=1}^r$ are the first r principal components of $\mathbb{X}_v$ with standardized predictors, s_t denotes the seasonal component of the monthly headline inflation extracted using X-13-Arima-Seats and p_f is selected using the Bic. As my intention here is to compare the linear and non-linear approach based on the same information set, I restrict the number of common factors entering this model to $r = 3$. I iterate this model forward to provide predictions of y_{t+h}.
7. The judgement forecast (JD). The forecasts are prepared by the professional forecasters surveyed on a monthly basis by *Bloomberg*. Commonly, the forecasts are prepared for the following quarters and not months. Therefore I use a Denton-Cholette temporal disaggregation (Dagum and Cholette, 2006) to obtain forecasts at the monthly frequency. For each vintage v I check the latest available information regarding the headline inflation at the time of preparing the forecast by the professional forecaster and use it as the first boundary value. The last boundary value is the forecast for the last quarter in the forecasting horizon. By using the Denton-Cholette method I disaggregate the quarterly data into monthly frequency. This approach is motivated by the findings of Faust and Wright (2013) stating that choosing two boundary values and interpolating between them is often found to have superior predictive ability. A visual inspection of these disaggregated series indicates that these are in fact smooth transitions between the initial and final boundary values.
8. The dynamic factor model (DFM). This model is described in Stelmasiak and Szafrański (2016b). Within this model the information set as in Baranowski et al. (2010) is reduced from a large number of variables to several unobserved factors. The model is estimated using the the two-step procedure of Doz et al. (2011). The forecast is obtained by iterating the model forward.
9. The Bayesian vector autoregressive model with the Sims-Zha priors (SZ). The specification of the model is described in Stelmasiak and Szafrański (2016a) and assumes that inflation is driven by a small number of variables. The model in the structural form contains twelve lags and includes deterministic seasonal dummies with loose priors. The forecast is obtained by iterating the model forward.
10. The Bayesian vector autoregressive model with the Villani steady-state prior (VI). The specification of the non-linear model is described in Stelmasiak and Szafrański (2016a). The model is a stationary reduced-form BVAR defined for deviations from the seasonal means with the very tight prior on the unconditional mean. The forecast is obtained by iterating the model forward.
11. The forecast combination of the bagged single hidden-layer feed-forward artificial neural networks described in this paper (ANN).
12. The combination of the two best performing models (CB1). The combination is derived after the monthly forecasts are annualized using equal weights.
13. A convolution of the two best performing models and the ANN model (CB2). The combination is derived after the monthly forecasts are annualized using equal weights.

With the exception of the judgement forecasts, all models are estimated on a monthly basis using the information set not larger than $\mathbb{X}_v$ available at the time of preparing the forecast. Since I conduct the evaluation of forecasts using the yearly CPI measure, I compute the annual inflation rate throughout the forecasting period based on the forecasts of the monthly headline inflation in the following manner[1]:

$$\hat{y}^{yy}_{t+h} = \begin{cases} \prod_{i=1}^{h}(1+\frac{\hat{y}_{t+i}}{100})\prod_{j=1}^{12-h}(1+\frac{y_{t-j+1}}{100}) & \text{if} \quad h<12 \\ \prod_{i=1}^{h}(1+\frac{\hat{y}_{t+i}}{100}) & \text{if} \quad h=12 \end{cases} \tag{20}$$

3.6. Forecast evaluation criteria

In the paper I perform two separate forecast comparisons. Firstly, in order to formally compare the accuracy of the point forecasts of the benchmark models and the artificial neural network's combined forecast, I use two most common evaluation criteria, namely the mean forecasting error (Mfe) and the mean square root forecasting error (Rmsfe). These statistics calculated for the yearly headline inflation are given as follows:

$$MFE_h = \frac{1}{K_h}\sum_{t=2010:12}^{2016:12}(y^{yy}_{t+h} - y^{yy}_{t+h|t}) \tag{21}$$

$$RMSFE_h = \left[\frac{1}{K_h}\sum_{t=2010:12}^{2016:12}(y^{yy}_{t+h} - y^{yy}_{t+h|t})^2\right]^{\frac{1}{2}} \tag{22}$$

where: K_h is the total number of out-of-sample inflation forecasts for a given horizon, h is the forecast horizon, $y^{yy}_{t+h|t}$ is the yearly headline inflation forecast of a given model at time t and y^{yy}_t is the realization. In my specific study K_h varies from 72 to 61, depending on the forecast horizon h.

In order to assess the statistical significance between the accuracy of the forecasts I perform the test of unconditional predictive accuracy (Giacomini and White, 2006). The null hypothesis states that the average loss from the competing models is equal $H_0 : E[L_{t+h}(e_{t+h,M_A})] - E[L_{t+h}(e_{t+h,M_B})] = 0$. I utilize a common quadratic loss function, hence the null hypothesis can be simplified to: $H_0 : e^2_{t+h,M_A} - e^2_{t+h,M_B} = 0$. The test statistic follows the Chi-squared distribution.

Secondly, I compare the quasi-distributions of different specifications of the artificial neural network models given by all point forecasts. For that purpose, I employ a kernel density estimation to approximate the density of all point forecasts for a given horizon. Following Groen et al. (2013) I use continuous ranked probability score – a strictly proper scoring rule that provides summary measure for the evaluation of probabilistic forecast (Gneiting and Raftery, 2007). Moreover, it is shown to reward forecast realization that are close to the middle of the forecast density. I resign to compare the logarithmic scores as they can be severely low for events with a very low probability, may take negative values and are more sensitive to outliers. The Crps is given by the following equation:

$$CRPS(F, y^{yy}_{t+h}) = \int_{-\infty}^{\infty}[F(y^{yy}_{t+h|t}) - \mathbb{1}_{\{y^{yy}_{t+h} \leq y^{yy}_{t+h|t}\}}]^2 dy^{yy}_{t+h} \tag{23}$$

where: $F(y^{yy}_{t+h|t})$ denotes the cumulative distribution function of a density forecast for $y^{yy}_{t+h,t}$ and y^{yy}_{t+h} denotes the realization of the forecast variable. I approximate the predictive density as well as the cumulative distribution function using the kernel distribution estimator. Due to the non-normal distribution of the quasi-densities, I apply numerical methods to calculate this measure.

3.7. Robustness check

In order to check whether the forecasting performance of the model is sensitive to the key assumptions regarding the parameters I perform an extensive robustness check and provide results for different

[1] As the weights in the CPI basket in Poland are revised on a yearly basis, I include minor corrections for the realized values used for annualizing the monthly CPI inflation in order to prevent making errors resulting from this aggregation method.

specifications of my approach. Since experimenting over the whole parameter space is computationally infeasible I estimate thirteen alternative specifications that differ from the baseline model with respect to the amount of information included in the training set δ_1, the complexity of the single neural network's architecture, i.e. the number of common factors treated as explanatory variables r and the expected extension of the hidden layer γ as well as the the key parameter of the learning algorithm Δ_{ij} and θ. Additionally, I also check whether the number of the generated bootstrap samples B influences the quality of the forecasts. In the limiting scenario bootstrap aggregating is not employed. As I insist on having 10 000 point forecasts for combination, decreasing B urges to increase the number of the neural networks D estimated for a given bootstrap pseudo-new learning set. Table 1 summarizes all considered specifications of my proposed forecasting approach. Throughout the paper, I treat specification I as the baseline model.

Table 1: The baseline and alternative specifications of the model

	Specification													
Parameter	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
δ_1	0,7	**0,6**	**0,8**	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7
r	3	3	3	**5**	**7-8 [*]**	3	3	3	3	3	3	3	3	3
γ	3	3	3	3	3	**6**	**9**	3	3	3	3	3	3	3
Δ_{ij}	0,1	0,1	0,1	0,1	0,1	0,1	0,1	**0,2**	0,1	0,1	0,1	0,1	0,1	0,1
θ	2	2	2	2	2	2	2	2	**1**	**3**	2	2	2	2
B	100	100	100	100	100	100	100	100	100	100	**50**	**25**	**10**	0
D	100	100	100	100	100	100	100	100	100	100	200	400	1000	**10000**

Note: The table presents different specifications of the model considered in the paper. Specification I is treated as the baseline model. Sensitivity analysis is conducted by comparing the inflation forecasts from the baseline model and the alternative specifications (II-XIV) and concerns the amount of information fed to the training sample and to the whole model, the average complexity of the hidden layer, the parameters of the resilient propagation algorithm as well as the extent and the necessity of employing bootstrap aggregation. For each specification the changed parameter is reported in the table in bold. Source: own calculations.

4. Results

In this section I present the results of the forecasting horse race conducted in the paper. Firstly, I assess the relative accuracy of all models using the whole out-of-sample period (2011:01-2016:12). For this purpose I present the RMSFE and the MFE statistics and report the outcomes of the Giacomini and White (2006) test of unconditional predictive ability. Secondly, in the analogous way I present the results for the excessive disinflation period (2012:01-2016:12). I single out this period deliberately due to the particular, uncommon behaviour of the headline inflation that has posed a demanding challenge for all forecasting models. I analyse the forecasting accuracy of the competing models and provide and explanation regarding the changes in their predictive accuracy in this period compared to the whole out-of-sample evaluation period. Lastly, I present the outcomes for the robustness analysis by presenting the errors statistics and the comparison of the ranked continuous probability scores in the two distinguished periods across considered specifications.

4.1. The full sample

I start the analysis of the results by reporting the RMSFE and the MFE statistics for all examined models. These measures are calculated using the whole out-of-sample evaluation period. In the second column of the table 2 I present the errors for the forecast combination of the bagged single hidden-layer feed-forward artificial neural network (ANN). In the next columns, I report the RMSFE statistics as the ratios of the RMSFE of the competing approach and the RSMFE of the ANN model. The ratio above one indicates that the competing model is characterized by lower forecast accuracy in comparison to the ANN. The biases of all examined models are included in the parentheses.

The general assessment of the ANN forecasts' quality is satisfactory. In the pairwise comparison the accuracy of the ANN forecast as measured by the RMSFE is higher in 95 out of 144 cases. Moreover, the bias of the forecast throughout the forecast horizon is only slightly negative and among the lowest for all examined models. To be precise, for horizon $h = 12$ it amounts to -0.2 pp. indicating that on average headline inflation in the whole period has been mildly overestimated.

The largest gain in the predictive accuracy is observed in comparison to the linear equivalent of the ANN model – the factor augmented vector autoregressive model estimated on the same information employed in the non-linear framework. Thus the only thing distinguishing these two approaches is the estimation method. As can be observed, the combination of random sampling, bootstrap aggregating and forecast pooling delivers substantial increase in the predictive accuracy. This considerable disproportion stems from the fact, that the FAV model is defined as a stationary process and as a results its forecast converges to the long-term mean throughout the forecast horizon. In the period of declining and persistently low headline inflation the overall accuracy of this model is poor and the forecast becomes severely biased, especially in the long-horizons. The ANN model on the other hand is not constrained by the stationarity requirement and produces forecast that do not account for the long-term mean but better predict the local price dynamics.

The ANN model outperforms also several other approaches. Not surprisingly, simple time-series models produce inferior forecasts. For more sophisticated approaches, the situation becomes more diverse. The DFM model becomes less accurate as the forecast horizon rises and for $h = 12$ it produces around 30% larger errors than the ANN model. There is also a weak evidence of more accurate forecasts for the short-term horizons of the ANN model in comparison to the VAR models estimated in a Bayesian fashion. Finally, the accuracy of the predictions prepared by the professional forecasters is quite low. Interestingly, these forecasts are also characterized by a substantial negative bias which can be explained by the anchoring of inflation expectations – a strong belief that the central bank will adjust the monetary policy to systematically bring back inflation towards its target.

There are, however, several cases for which the forecast from the ANN model is less accurate. Firstly, the DFM model outperforms the ANN model by a large margin (around 14%) for the nowcast of the headline inflation ($h = 1$) and remains more accurate for horizons $h \leq 6$. Secondly, the small VARs with a long lag polynomial estimated using Bayesian methods improve upon the ANN forecast quite substantially for longer horizons. In particular, the RMSFE for the VI model with a highly restrictive prior on the steady state is lower by around 10% for horizons $h > 5$. Finally, the equal weight forecast combination of the two best performing alternative models (CB1) delivers significant improvement in the forecast accuracy throughout the whole forecast horizon, even though they are both slightly worse for horizon $h = 3$ than the ANN model. By combining the linear and non-linear BVAR model the forecast accuracy is greater, especially in the longer-term (the decrease in the RMSFE amounts to around 10-15% for forecast horizon $h \geq 6$). The combination forecast also yields greater accuracy for horizons $h = \{3, 4\}$, for which both models considered separately present inferior accuracy.

The superior predictive performance of the CB1 model can undermine the mostly positive results. However, to every cloud there is a silver lining. By combining further the CB1 and the ANN forecasts, the accuracy rises still for all considered horizons. This indicates that by pooling forecasts from various linear and non-linear approaches one can achieve further gains in the predictive accuracy.

An important question remains whether the relative improvements in the forecast accuracy of the examined models are statistically significant. To address this issue I employ the Giacomini and White (2006) test of unconditional superior accuracy. Following a common approach, I utilize a quadratic loss function, hence the test hypothesis can be simplified to $e^2_{t+h,M_A} - e^2_{t+h,M_B} = 0$. I denote here the forecast from the artificial neural networks as M_A and the forecast from the competing model as M_B.

Table 3 presents the test statistic, its sign in parentheses (indicating if model M_A generates larger or smaller average loss denoted as '+' or '−', respectively), and the p-value corresponding to the calculated test statistic (in square brackets). To facilitate the analysis of the reported test results, I report the cases where the ANN model is statistically better (worse) than the forecast from the competing model in bold (italics). The significance is checked at $\alpha = 0.1$ significance level.

On the one hand, among the 95 cases, in which the ANN model beats the competing model in terms of the forecast accuracy, 44 turn out to be statistically significant. The ANN model significantly outperforms the standard RW for all horizons and the AR1 as well as the FAV model for horizons $h \geq 2$, but for the nowcast the forecasts appear to have comparable accuracy. It appears also that the judgement forecast is significantly worse both for short and longer-term. Lastly, the ANN model is statistically more accurate than the DFM model only for $h = 12$.

On the other hand, the forecast accuracy of the ANN model is significantly inferior only in 9 cases. These include the nowcast of the DFM model (probably due to the fact, that the DFM model includes the expectation maximization algorithm for missing data, whereas the ragged edges in this paper are simply vertically aligned) and the CB1 model. Moreover, the combination of the three best performing linear and non-linear models prove more accurate for horizons $h = \{5, \ldots, 11\}$ and only marginally insignificant for $h = 12$. This is an interesting conclusion indicating that in order to deliver statistically superior forecasts one should consider several linear and non-linear approaches with various forecast accuracy over the forecast horizon.

The results for the whole sample can be perceived overall as satisfactory once the test for the difference of the average loss is applied. The combination of the bagged single hidden-layer feed-forward artificial neural networks provides an improvement in the forecast accuracy in the majority of cases and generally is characterized by a relatively small forecast bias. Moreover, by combining several different approaches, further statistically significant gains in the forecast accuracy are observed. However, two flies in the ointment need a short commentary. First, in the majority of cases (91 to be precise) the predictive ability of the proposed and competing model is comparable despite the fact that in the majority of cases the RMSFE of the ANN is lower in comparison to the competing approaches. Second, the non-linear model fails to beat the Atkeson and Ohanian (2001) specification of the random walk, though it provides around 20% increase in the accuracy as measured by the RMSFE at $h = 12$ and generally is characterized by a

Table 2: The root mean square forecasting errors and mean forecasting errors of the competing models in the whole evaluation period

	Specification												
Horizon	ANN	RW	AO	AR1	AR12	BS	DFM	FAV	SZ	VI	CB1	CB2	JD
h=1	0,286 (0,006)	1,268 (-0,014)	1,139 (-0,024)	1,191 (-0,091)	1,070 (-0,065)	1,057 (-0,029)	0,858 (-0,035)	1,076 (-0,078)	0,921 (-0,008)	1,048 (-0,033)	0,929 (-0,020)	0,896 (-0,012)	1,327 (-0,154)
h=2	0,422 (0,006)	1,406 (-0,037)	1,232 (-0,060)	1,320 (-0,237)	1,113 (-0,159)	1,060 (-0,069)	0,901 (-0,066)	1,233 (-0,219)	0,991 (-0,017)	1,061 (-0,051)	0,956 (-0,034)	0,896 (-0,021)	1,279 (-0,269)
h=3	0,600 (0,007)	1,430 (-0,055)	1,172 (-0,100)	1,334 (-0,408)	1,075 (-0,255)	1,004 (-0,106)	0,880 (-0,124)	1,285 (-0,382)	1,003 (-0,036)	1,019 (-0,068)	0,946 (-0,052)	0,894 (-0,033)	1,205 (-0,396)
h=4	0,720 (-0,011)	1,536 (-0,096)	1,151 (-0,162)	1,424 (-0,616)	1,100 (-0,370)	1,020 (-0,172)	0,907 (-0,204)	1,381 (-0,588)	1,039 (-0,070)	1,007 (-0,090)	0,957 (-0,080)	0,904 (-0,057)	1,260 (-0,551)
h=5	0,847 (-0,028)	1,620 (-0,147)	1,082 (-0,236)	1,475 (-0,842)	1,082 (-0,503)	0,999 (-0,249)	0,949 (-0,301)	1,435 (-0,806)	1,012 (-0,112)	0,955 (-0,116)	0,912 (-0,114)	0,878 (-0,085)	1,352 (-0,711)
h=6	0,937 (-0,053)	1,733 (-0,221)	1,037 (-0,324)	1,562 (-1,086)	1,081 (-0,643)	0,989 (-0,339)	0,965 (-0,419)	1,514 (-1,038)	0,969 (-0,168)	0,908 (-0,139)	0,858 (-0,154)	0,838 (-0,120)	1,620 (-0,877)
h=7	1,006 (-0,069)	1,855 (-0,271)	1,056 (-0,404)	1,680 (-1,319)	1,135 (-0,773)	1,037 (-0,410)	1,003 (-0,547)	1,633 (-1,253)	0,963 (-0,216)	0,901 (-0,137)	0,853 (-0,176)	0,830 (-0,141)	1,742 (-0,997)
h=8	1,083 (-0,085)	1,955 (-0,317)	1,071 (-0,484)	1,782 (-1,549)	1,185 (-0,897)	1,080 (-0,491)	1,027 (-0,681)	1,732 (-1,464)	0,947 (-0,265)	0,887 (-0,115)	0,846 (-0,190)	0,821 (-0,155)	1,671 (-1,101)
h=9	1,143 (-0,105)	2,082 (-0,378)	1,099 (-0,568)	1,909 (-1,781)	1,259 (-1,029)	1,142 (-0,567)	1,105 (-0,822)	1,849 (-1,675)	0,972 (-0,314)	0,907 (-0,086)	0,877 (-0,200)	0,839 (-0,168)	1,545 (-1,191)
h=10	1,196 (-0,120)	2,167 (-0,450)	1,125 (-0,665)	2,030 (-2,023)	1,323 (-1,168)	1,182 (-0,672)	1,167 (-0,979)	1,965 (-1,894)	0,985 (-0,373)	0,909 (-0,068)	0,887 (-0,220)	0,840 (-0,187)	1,564 (-1,299)
h=11	1,259 (-0,157)	2,263 (-0,555)	1,148 (-0,788)	2,138 (-2,288)	1,371 (-1,327)	1,194 (-0,789)	1,227 (-1,162)	2,061 (-2,141)	0,964 (-0,458)	0,896 (-0,051)	0,882 (-0,255)	0,831 (-0,222)	1,580 (-1,417)
h=12	1,310 (-0,204)	2,239 (-0,757)	1,193 (-0,923)	2,260 (-2,571)	1,428 (-1,501)	1,237 (-0,932)	1,296 (-1,364)	2,166 (-2,398)	0,962 (-0,547)	0,890 (-0,039)	0,884 (-0,293)	0,826 (-0,263)	1,596 (-1,535)

Note: The table presents the root mean squared forecasting errors and the mean forecasting errors (in parentheses) for the competing models in the whole out-of-sample evaluation period (2011:01-2016:12). For the combination forecast of the bagged single hidden-layer feed-forward artificial neural networks the error statistics are reported in levels. For the other competing models the RMSFE is presented as a ratio of the RMSFE of the competing and the ANN model. A ratio greater than one indicates that the competing model produces on average less accurate forecasts. Source: own calculations.

Table 3: Results of the Giacomini and White (2006) test of unconditional predictive accuracy in the whole evaluation period

	Specification											
Horizon	ANN vs RW	ANN vs AO	ANN vs AR1	ANN vs AR12	ANN vs BS	ANN vs DFM	ANN vs FAV	ANN vs SZ	ANN vs VI	ANN vs CB1	ANN vs CB2	ANN vs JD
h=1	**2,755 (-) [0,097]**	1,802 (-) [0,180]	2,211 (-) [0,137]	0,602 (-) [0,438]	0,513 (-) [0,474]	*4,596 (+) [0,032]*	0,500 (-) [0,479]	0,644 (+) [0,422]	0,210 (-) [0,647]	*0,547 (+) [0,460]*	2,179 (+) [0,140]	**7,196 (-) [0,007]**
h=2	**4,760 (-) [0,029]**	2,338 (-) [0,126]	**4,675 (-) [0,031]**	1,223 (-) [0,269]	0,406 (-) [0,524]	1,652 (+) [0,199]	**3,473 (-) [0,062]**	0,006 (+) [0,939]	0,294 (-) [0,587]	0,160 (+) [0,689]	1,720 (+) [0,190]	**3,836 (-) [0,050]**
h=3	**6,562 (-) [0,010]**	1,454 (-) [0,228]	**6,180 (-) [0,013]**	0,646 (-) [0,422]	0,004 (-) [0,952]	1,800 (+) [0,180]	**5,611 (-) [0,018]**	0,001 (-) [0,980]	0,025 (-) [0,874]	0,217 (+) [0,641]	1,594 (+) [0,207]	1,550 (-) [0,213]
h=4	**8,231 (-) [0,004]**	1,352 (-) [0,245]	**9,555 (-) [0,002]**	1,154 (-) [0,283]	0,083 (-) [0,773]	1,449 (+) [0,229]	**10,757 (-) [0,001]**	0,124 (-) [0,725]	0,005 (-) [0,942]	0,178 (+) [0,673]	1,770 (+) [0,183]	1,920 (-) [0,166]
h=5	**10,736 (-) [0,001]**	0,470 (-) [0,493]	**8,939 (-) [0,003]**	0,501 (-) [0,479]	0,000 (+) [0,987]	0,447 (+) [0,504]	**10,659 (-) [0,001]**	0,018 (-) [0,893]	0,188 (+) [0,665]	0,822 (+) [0,365]	*3,020 (+) [0,082]*	2,094 (-) [0,148]
h=6	**12,338 (-) [0,000]**	0,079 (-) [0,778]	**7,989 (-) [0,005]**	0,289 (-) [0,591]	0,012 (+) [0,914]	0,159 (+) [0,690]	**8,417 (-) [0,004]**	0,112 (+) [0,738]	0,539 (+) [0,463]	1,569 (+) [0,210]	*3,747 (+) [0,053]*	2,165 (-) [0,141]
h=7	**12,333 (-) [0,000]**	0,157 (-) [0,692]	**8,430 (-) [0,004]**	0,579 (-) [0,447]	0,106 (-) [0,745]	0,001 (-) [0,974]	**8,414 (-) [0,004]**	0,137 (+) [0,711]	0,513 (+) [0,474]	1,423 (+) [0,233]	*3,539 (+) [0,060]*	2,537 (-) [0,111]
h=8	**13,711 (-) [0,000]**	0,258 (-) [0,612]	**8,962 (-) [0,003]**	0,966 (-) [0,326]	0,445 (-) [0,505]	0,065 (-) [0,798]	**8,565 (-) [0,003]**	0,261 (+) [0,610]	0,627 (+) [0,429]	1,537 (+) [0,215]	*3,921 (+) [0,048]*	**3,418 (-) [0,064]**
h=9	**14,316 (-) [0,000]**	0,509 (-) [0,476]	**9,504 (-) [0,002]**	1,783 (-) [0,182]	1,541 (-) [0,214]	0,834 (-) [0,361]	**8,632 (-) [0,003]**	0,065 (+) [0,798]	0,414 (+) [0,520]	1,085 (+) [0,298]	*3,717 (+) [0,054]*	**3,923 (-) [0,048]**
h=10	**14,377 (-) [0,000]**	0,612 (-) [0,434]	**9,218 (-) [0,002]**	2,367 (-) [0,124]	1,805 (-) [0,179]	1,647 (-) [0,199]	**8,512 (-) [0,004]**	0,015 (+) [0,901]	0,398 (+) [0,528]	0,864 (+) [0,353]	*3,502 (+) [0,061]*	**3,792 (-) [0,051]**
h=11	**14,876 (-) [0,000]**	0,635 (-) [0,426]	**8,746 (-) [0,003]**	**2,772 (-) [0,096]**	1,630 (-) [0,202]	2,269 (-) [0,132]	**7,936 (-) [0,005]**	0,052 (+) [0,820]	0,542 (+) [0,462]	0,795 (+) [0,373]	*3,283 (+) [0,070]*	**3,382 (-) [0,066]**
h=12	**11,261 (-) [0,001]**	0,962 (-) [0,327]	**8,327 (-) [0,004]**	**3,480 (-) [0,062]**	2,066 (-) [0,151]	**3,000 (-) [0,083]**	**7,364 (-) [0,007]**	0,037 (+) [0,847]	0,540 (+) [0,462]	0,554 (+) [0,457]	2,563 (+) [0,109]	**2,892 (-) [0,089]**

Note: The table presents the results of the Giacomini and White (2006) test of unconditional predictive accuracy calculated in the whole out-of-sample evaluation period (2011:01-2016:12). Under the null $H_0 : e^2_{t+h,M_A} - e^2_{t+h,M_B} = 0$. M_A, M_B *and* e_{t+h} *denote the artificial neural network model, the competing model and the forecast error for the horizon* h*, respectively. In the table I report the test statistics, the sign (in parentheses) indicating, whether model* M_A *delivers more accurate (as denoted by '–') or less accurate (as denoted by '+') forecasts and the p-value corresponding to the test statistic (in square brackets). Statistically more accurate forecast from the model* M_A *and* M_B *are reported in bold and italics, respectively. Source: own calculations.*

much smaller bias. This result corroborates the thesis that the majority of more sophisticated models fails to improve in a statistically significant manner upon this simple benchmark.

In the next section I address the question whether the conclusions change after the analysis of the out-of-sample forecasts evaluation is restricted to the period of systematically falling and persistently low inflation.

4.2. The excessive disinflation period

The price development in Poland after 2011 has followed an unorthodox – yet common for the majority of the economies during this time period – phenomenon of excessive disinflation (Constâncio, 2015). Starting in the first quarter of 2012, the headline inflation has abruptly fallen from 4.3% y-o-y in January 2012 to -1.6% in January 2016. It has remained persistently low and negative throughout the 2016, which is baffling considering the developments in the domestic real activity. Since then headline inflation has quite vigorously rebounded driven by the developments in the external factors.

The peculiar behaviour of headline inflation in this time period has created a difficult challenge for both forecasting models and professional forecasters alike. On the one hand, most of the traditional, stationary models has produced forecasts converging to the unconditional mean throughout the forecast horizon. On the other hand, the forecasts prepared by financial analysts revealed a strong belief that the central bank will eventually adjust its monetary policy to bring inflation gradually back to its desired target. As a result, both approaches overestimated inflation quite considerably for a long period of time.

In this section I evaluate the accuracy of the competing models in this slightly restricted out-of-sample time period (2012:01-2016:12). In particular, I want to examine whether the non-linear approaches are able to better approximate the inflation developments in this period.

Table 4 presents the RMSFE and the MFE of the examined models in an analogous fashion to the table 2. Additionally, for this out-of-sample period I also assess the statistical significance of the unconditional predictive accuracy of the models by utilizing the same specification of the Giacomini and White (2006) test. Table 5 presents the results of this procedure in an analogous way to the table 3.

In the pairwise comparison of the point forecast accuracy, the ANN model outperforms a competing benchmark in 95 cases, similarly as in the case of the whole out-of-sample evaluation period. The accuracy of the ANN model in this sample slightly rises, which is opposite for most of the other competing models. This is an important conclusion – a structural change in the development of prices in Poland has not affected the quality of the forecasts, most probably due to the employed thick modelling framework. As a result, the Giacomini and White (2006) test statistics points to the superiority of the forecasts generated by the ANN model in 56 pairwise comparisons (increase from 44 in the whole out-of-sample evaluation period). Across the examined models, the superior predictive ability is observed mainly for longer horizons. The forecast from this model is consistently better than the RW, AR1, FAV and JD forecasts.

The distinguished outliers are again the SZ and VI models estimated using the Bayesian methods. Both these frameworks are able to increase the forecast accuracy of the ANN forecast for the majority of examined horizons. Moreover, the VI model is statistically more accurate than the ANN model for horizons $h = \{7, 8, 9, 10\}$ which is an aftermath of the very tight prior imposed on the unconditional mean. As a result, the model accounts for the slowly evolving local mean of inflation.

Finally, the increase in the forecast accuracy of the CB1 model amounts to around 22%, but is insufficient to be statistically significant. However, once the forecasts from the SZ, the VI and the ANN model are pooled together, the increase in the accuracy of the combined forecasts is statistically significant for horizons $h \leq 10$. This corroborates the previous finding stating that a careful combination of models diversified both in terms of employed methods as well as quality across horizons should provide further gains in accuracy in comparison to the individual models.

Table 4: The root mean square forecasting errors and mean forecasting errors of the competing models in the excessive disinflation period

	Specification												
Horizon	ANN	RW	AO	AR1	AR12	BS	DFM	FAV	SZ	VI	CB1	CB2	JD
h=1	0,256 (0,000)	1,253 (-0,011)	1,101 (-0,034)	1,194 (-0,111)	1,003 (-0,084)	1,006 (-0,037)	0,913 (-0,031)	1,011 (-0,100)	0,826 (-0,016)	1,031 (-0,061)	0,851 (-0,038)	0,825 (-0,025)	1,462 (-0,227)
h=2	0,392 (-0,001)	1,412 (-0,028)	1,161 (-0,079)	1,355 (-0,287)	1,079 (-0,203)	1,044 (-0,083)	0,930 (-0,056)	1,228 (-0,276)	0,886 (-0,039)	1,006 (-0,111)	0,856 (-0,075)	0,815 (-0,050)	1,402 (-0,376)
h=3	0,551 (-0,004)	1,422 (-0,043)	1,095 (-0,133)	1,392 (-0,503)	1,071 (-0,329)	1,010 (-0,136)	0,898 (-0,123)	1,309 (-0,488)	0,894 (-0,081)	0,949 (-0,165)	0,834 (-0,123)	0,802 (-0,083)	1,327 (-0,546)
h=4	0,654 (-0,025)	1,524 (-0,082)	1,090 (-0,210)	1,535 (-0,762)	1,141 (-0,472)	1,064 (-0,216)	0,941 (-0,213)	1,466 (-0,749)	0,940 (-0,136)	0,927 (-0,223)	0,841 (-0,180)	0,810 (-0,128)	1,444 (-0,743)
h=5	0,764 (-0,044)	1,611 (-0,134)	1,053 (-0,304)	1,639 (-1,047)	1,172 (-0,636)	1,061 (-0,312)	1,027 (-0,326)	1,565 (-1,030)	0,959 (-0,202)	0,874 (-0,292)	0,816 (-0,247)	0,799 (-0,179)	1,600 (-0,967)
h=6	0,862 (-0,059)	1,668 (-0,200)	1,050 (-0,402)	1,750 (-1,343)	1,202 (-0,797)	1,054 (-0,411)	1,048 (-0,456)	1,673 (-1,316)	0,952 (-0,276)	0,821 (-0,347)	0,781 (-0,312)	0,777 (-0,227)	1,901 (-1,207)
h=7	0,947 (-0,078)	1,727 (-0,244)	1,101 (-0,499)	1,877 (-1,637)	1,273 (-0,958)	1,121 (-0,498)	1,096 (-0,611)	1,802 (-1,591)	0,965 (-0,354)	0,794 (-0,384)	0,785 (-0,369)	0,779 (-0,272)	1,992 (-1,417)
h=8	1,033 (-0,099)	1,802 (-0,284)	1,147 (-0,597)	1,998 (-1,929)	1,333 (-1,113)	1,171 (-0,601)	1,127 (-0,778)	1,922 (-1,866)	0,965 (-0,437)	0,749 (-0,399)	0,780 (-0,418)	0,776 (-0,312)	1,874 (-1,585)
h=9	1,111 (-0,128)	1,890 (-0,344)	1,182 (-0,705)	2,126 (-2,230)	1,398 (-1,283)	1,232 (-0,700)	1,207 (-0,961)	2,043 (-2,145)	0,987 (-0,527)	0,740 (-0,414)	0,801 (-0,471)	0,794 (-0,356)	1,698 (-1,708)
h=10	1,164 (-0,148)	1,994 (-0,414)	1,230 (-0,824)	2,286 (-2,541)	1,480 (-1,459)	1,283 (-0,832)	1,274 (-1,164)	2,196 (-2,433)	1,002 (-0,629)	0,721 (-0,441)	0,807 (-0,535)	0,797 (-0,406)	1,724 (-1,855)
h=11	1,215 (-0,179)	2,115 (-0,509)	1,281 (-0,961)	2,443 (-2,868)	1,553 (-1,648)	1,316 (-0,957)	1,354 (-1,389)	2,343 (-2,742)	0,993 (-0,753)	0,703 (-0,458)	0,813 (-0,606)	0,800 (-0,464)	1,764 (-2,009)
h=12	1,269 (-0,219)	2,053 (-0,716)	1,335 (-1,102)	2,590 (-3,210)	1,618 (-1,839)	1,359 (-1,096)	1,428 (-1,632)	2,475 (-3,055)	0,993 (-0,874)	0,700 (-0,473)	0,824 (-0,673)	0,805 (-0,522)	1,788 (-2,154)

Note: The table presents the root mean squared forecasting errors and the mean forecasting errors (in parentheses) for the competing models during the excessive disinflation period (2012:01-2016:12). For the combination forecast of the bagged single hidden-layer feed-forward artificial neural networks the error statistics are reported in levels. For the other competing models the RMSFE is presented as the ratio of the RMSFE of the competing and the ANN model. A ratio greater than one indicates that the competing model produces on average less accurate forecasts. Source: own calculations.

Table 5: Results of the Giacomini and White (2006) test of unconditional predictive accuracy in the excessive disinflation evaluation period

	Specification											
Horizon	ANN vs RW	ANN vs AO	ANN vs AR1	ANN vs AR12	ANN vs BS	ANN vs DFM	ANN vs FAV	ANN vs SZ	ANN vs VI	ANN vs CB1	ANN vs CB2	ANN vs JD
h=1	1,237 (-) [0,266]	0,506 (-) [0,477]	1,104 (-) [0,293]	0,001 (-) [0,978]	0,004 (-) [0,949]	1,135 (+) [0,287]	0,006 (-) [0,939]	1,727 (+) [0,189]	0,042 (-) [0,838]	1,232 (+) [0,267]	*3,040 (+) [0,081]*	**9,886 (-) [0,002]**
h=2	**2,807 (-) [0,094]**	0,725 (-) [0,394]	**3,304 (-) [0,069]**	0,403 (-) [0,526]	0,128 (-) [0,721]	0,607 (+) [0,436]	1,958 (-) [0,162]	0,642 (+) [0,423]	0,002 (-) [0,965]	1,089 (+) [0,297]	*3,244 (+) [0,072]*	**5,574 (-) [0,018]**
h=3	**3,590 (-) [0,058]**	0,279 (-) [0,598]	**4,600 (-) [0,032]**	0,308 (-) [0,579]	0,010 (-) [0,919]	0,707 (+) [0,400]	**3,468 (-) [0,063]**	0,455 (+) [0,500]	0,102 (+) [0,749]	1,096 (+) [0,295]	*2,862 (+) [0,091]*	2,463 (-) [0,117]
h=4	**4,607 (-) [0,032]**	0,261 (-) [0,609]	**8,063 (-) [0,005]**	1,171 (-) [0,279]	0,477 (-) [0,490]	0,331 (+) [0,565]	**8,267 (-) [0,004]**	0,192 (+) [0,661]	0,329 (+) [0,566]	1,410 (+) [0,235]	*3,857 (+) [0,050]*	**3,851 (-) [0,050]**
h=5	**6,420 (-) [0,011]**	0,090 (-) [0,764]	**9,121 (-) [0,003]**	1,210 (-) [0,271]	0,332 (-) [0,564]	0,086 (-) [0,769]	**9,580 (-) [0,002]**	0,110 (+) [0,740]	1,206 (+) [0,272]	2,011 (+) [0,156]	*4,320 (+) [0,038]*	**4,293 (-) [0,038]**
h=6	**7,290 (-) [0,007]**	0,075 (-) [0,784]	**9,519 (-) [0,002]**	1,207 (-) [0,272]	0,175 (-) [0,676]	0,221 (-) [0,638]	**9,181 (-) [0,002]**	0,141 (+) [0,708]	2,287 (+) [0,130]	2,405 (+) [0,121]	*4,384 (+) [0,036]*	**3,093 (-) [0,079]**
h=7	**7,989 (-) [0,005]**	0,312 (-) [0,576]	**10,809 (-) [0,001]**	1,851 (-) [0,174]	0,839 (-) [0,360]	0,786 (-) [0,375]	**9,758 (-) [0,002]**	0,074 (+) [0,786]	*2,847 (+) [0,092]*	2,272 (+) [0,132]	*4,365 (+) [0,037]*	**3,183 (-) [0,074]**
h=8	**9,775 (-) [0,002]**	0,764 (-) [0,382]	**13,408 (-) [0,000]**	**2,730 (-) [0,099]**	1,570 (-) [0,210]	1,335 (-) [0,248]	**11,322 (-) [0,001]**	0,075 (+) [0,784]	*3,641 (+) [0,056]*	2,354 (+) [0,125]	*4,659 (+) [0,031]*	**4,347 (-) [0,037]**
h=9	**10,702 (-) [0,001]**	1,309 (-) [0,253]	**16,142 (-) [0,000]**	**4,003 (-) [0,045]**	**3,620 (-) [0,057]**	**3,468 (-) [0,063]**	**12,686 (-) [0,000]**	0,010 (+) [0,922]	*3,183 (+) [0,074]*	1,849 (+) [0,174]	*4,092 (+) [0,043]*	**5,556 (-) [0,018]**
h=10	**9,880 (-) [0,002]**	1,683 (-) [0,194]	**19,075 (-) [0,000]**	**5,330 (-) [0,021]**	**3,688 (-) [0,055]**	**4,854 (-) [0,028]**	**14,896 (-) [0,000]**	0,000 (-) [0,990]	*3,074 (+) [0,080]*	1,427 (+) [0,232]	*3,240 (+) [0,072]*	**5,676 (-) [0,017]**
h=11	**10,228 (-) [0,001]**	1,935 (-) [0,164]	**21,509 (-) [0,000]**	**6,876 (-) [0,009]**	**3,814 (-) [0,051]**	**6,141 (-) [0,013]**	**16,788 (-) [0,000]**	0,001 (+) [0,971]	2,674 (+) [0,102]	0,994 (+) [0,319]	2,400 (+) [0,121]	**5,518 (-) [0,019]**
h=12	**6,755 (-) [0,009]**	2,507 (-) [0,113]	**22,158 (-) [0,000]**	**8,823 (-) [0,003]**	**3,810 (-) [0,051]**	**6,783 (-) [0,009]**	**17,503 (-) [0,000]**	0,001 (+) [0,978]	2,023 (+) [0,155]	0,641 (+) [0,423]	1,676 (+) [0,195]	**4,902 (-) [0,027]**

Note: The table presents the results of the Giacomini and White (2006) test of unconditional predictive accuracy calculated during the excessive disinflation period (2012:01-2016:12). Under the null $H_0 : e^2_{t+h,M_A} - e^2_{t+h,M_B} = 0$. M_A, M_B and e_{t+h} denote the artificial neural network model, the competing model and the forecast error for horizon h, respectively. In the table I report the test statistics, the sign (in parentheses) indicating, whether model M_A delivers more accurate (as denoted by '–') or less accurate (as denoted by '+') forecasts and the p-value corresponding to the test statistic (in square brackets). Statistically more accurate forecast from the model M_A and M_B are reported in bold and italics, respectively. Source: own calculations.

Table 6: Results of the Giacomini and White (2006) test of unconditional predictive accuracy in the whole out-of-sample evaluation period

Horizon	CB2 vs RW	CB2 vs AO	CB2 vs AR1	CB2 vs AR12	CB2 vs BS	Specification CB2 vs DFM	CB2 vs FAV	CB2 vs SZ	CB2 vs VI	CB2 vs CB1	CB2 vs ANN	CB2 vs JD
h=1	**6,456 (-) [0,011]**	**8,629 (-) [0,003]**	**7,917 (-) [0,005]**	**8,623 (-) [0,003]**	**6,239 (-) [0,012]**	0,320 (+) [0,572]	**5,814 (-) [0,016]**	0,339 (-) [0,560]	**6,241 (-) [0,012]**	1,097 (-) [0,295]	2,179 (-) [0,140]	**11,349 (-) [0,001]**
h=2	**8,589 (-) [0,003]**	**7,355 (-) [0,007]**	**11,764 (-) [0,001]**	**10,421 (-) [0,001]**	**4,700 (-) [0,030]**	0,003 (-) [0,956]	**14,356 (-) [0,000]**	**2,909 (-) [0,088]**	**6,570 (-) [0,010]**	2,600 (-) [0,107]	1,720 (-) [0,190]	**8,659 (-) [0,003]**
h=3	**11,117 (-) [0,001]**	**5,490 (-) [0,019]**	**16,050 (-) [0,000]**	**6,336 (-) [0,012]**	2,640 (-) [0,104]	0,038 (+) [0,845]	**18,493 (-) [0,000]**	**3,499 (-) [0,061]**	**4,193 (-) [0,041]**	1,752 (-) [0,186]	1,594 (-) [0,207]	**4,803 (-) [0,028]**
h=4	**10,983 (-) [0,001]**	**4,894 (-) [0,027]**	**16,514 (-) [0,000]**	**4,266 (-) [0,039]**	1,798 (-) [0,180]	0,001 (-) [0,973]	**22,143 (-) [0,000]**	**4,772 (-) [0,029]**	**2,710 (-) [0,100]**	1,934 (-) [0,164]	1,770 (-) [0,183]	**3,897 (-) [0,048]**
h=5	**14,057 (-) [0,000]**	**3,445 (-) [0,063]**	**16,451 (-) [0,000]**	**2,823 (-) [0,093]**	1,651 (-) [0,199]	0,669 (-) [0,413]	**23,156 (-) [0,000]**	**5,974 (-) [0,015]**	1,372 (-) [0,242]	1,045 (-) [0,307]	**3,020 (-) [0,082]**	**3,808 (-) [0,051]**
h=6	**15,938 (-) [0,000]**	2,609 (-) [0,106]	**16,235 (-) [0,000]**	2,538 (-) [0,111]	2,134 (-) [0,144]	1,895 (-) [0,169]	**19,783 (-) [0,000]**	**5,597 (-) [0,018]**	0,806 (-) [0,369]	0,301 (-) [0,583]	**3,747 (-) [0,053]**	**3,210 (-) [0,073]**
h=7	**15,226 (-) [0,000]**	**2,727 (-) [0,099]**	**15,993 (-) [0,000]**	**2,921 (-) [0,087]**	**2,969 (-) [0,085]**	**3,261 (-) [0,071]**	**17,500 (-) [0,000]**	**4,452 (-) [0,035]**	0,769 (-) [0,381]	0,334 (-) [0,563]	**3,539 (-) [0,060]**	**3,559 (-) [0,059]**
h=8	**16,282 (-) [0,000]**	**2,861 (-) [0,091]**	**15,480 (-) [0,000]**	**3,570 (-) [0,059]**	**3,702 (-) [0,054]**	**4,567 (-) [0,033]**	**15,799 (-) [0,000]**	**3,217 (-) [0,073]**	0,599 (-) [0,439]	0,364 (-) [0,546]	**3,921 (-) [0,048]**	**5,338 (-) [0,021]**
h=9	**16,226 (-) [0,000]**	2,634 (-) [0,105]	**14,422 (-) [0,000]**	**4,413 (-) [0,036]**	**4,952 (-) [0,026]**	**6,564 (-) [0,010]**	**13,807 (-) [0,000]**	**3,051 (-) [0,081]**	0,552 (-) [0,458]	0,780 (-) [0,377]	**3,717 (-) [0,054]**	**7,396 (-) [0,007]**
h=10	**15,883 (-) [0,000]**	2,315 (-) [0,128]	**13,140 (-) [0,000]**	**5,288 (-) [0,021]**	**4,725 (-) [0,030]**	**8,598 (-) [0,003]**	**12,681 (-) [0,000]**	**3,026 (-) [0,082]**	0,466 (-) [0,495]	0,966 (-) [0,326]	**3,502 (-) [0,061]**	**7,290 (-) [0,007]**
h=11	**15,980 (-) [0,000]**	2,098 (-) [0,148]	**12,423 (-) [0,000]**	**6,165 (-) [0,013]**	**4,118 (-) [0,042]**	**10,634 (-) [0,001]**	**11,703 (-) [0,001]**	2,037 (-) [0,154]	0,375 (-) [0,540]	0,906 (-) [0,341]	**3,283 (-) [0,070]**	**6,905 (-) [0,009]**
h=12	**12,901 (-) [0,000]**	2,279 (-) [0,131]	**11,870 (-) [0,001]**	**7,422 (-) [0,006]**	**3,841 (-) [0,050]**	**12,468 (-) [0,000]**	**10,875 (-) [0,001]**	1,522 (-) [0,217]	0,334 (-) [0,563]	0,861 (-) [0,354]	2,563 (-) [0,109]	**6,314 (-) [0,012]**

Note: The table presents the results of the Giacomini and White (2006) test of unconditional predictive accuracy calculated during the whole out-of-sample evaluation period (2011:01-2016:12). Under the null $H_0 : e^2_{t+h,M_A} - e^2_{t+h,M_B} = 0$. M_A, M_B and e_{t+h} denote the CB2 model, the competing model and the forecast error for horizon h, respectively. In the table I report the test statistics, the sign (in parentheses) indicating, whether model M_A delivers more accurate (as denoted by '−') or less accurate (as denoted by '+') forecasts and the p-value corresponding to the test statistic (in square brackets). Statistically more accurate forecast from the model M_A and M_B are reported in bold and italics, respectively. Source: own calculations.

Table 7: Results of the Giacomini and White (2006) test of unconditional predictive accuracy in the excessive disinflation evaluation period

Horizon	CB2 vs RW	CB2 vs AO	CB2 vs AR1	CB2 vs AR12	CB2 vs BS	Specification CB2 vs DFM	CB2 vs FAV	CB2 vs SZ	CB2 vs VI	CB2 vs CB1	CB2 vs ANN	CB2 vs JD
h=1	**3,986 (-) [0,046]**	**6,184 (-) [0,013]**	**5,445 (-) [0,020]**	**5,789 (-) [0,016]**	**5,110 (-) [0,024]**	1,398 (-) [0,237]	**4,729 (-) [0,030]**	0,000 (-) [0,984]	**5,048 (-) [0,025]**	0,369 (-) [0,543]	**3,040 (-) [0,081]**	**17,472 (-) [0,000]**
h=2	**6,057 (-) [0,014]**	**4,684 (-) [0,030]**	**10,013 (-) [0,002]**	**8,691 (-) [0,003]**	**4,823 (-) [0,028]**	2,172 (-) [0,141]	**11,037 (-) [0,001]**	1,250 (-) [0,264]	**4,522 (-) [0,033]**	0,741 (-) [0,389]	**3,244 (-) [0,072]**	**15,489 (-) [0,000]**
h=3	**8,176 (-) [0,004]**	**3,754 (-) [0,053]**	**16,153 (-) [0,000]**	**7,302 (-) [0,007]**	**6,768 (-) [0,009]**	1,854 (-) [0,173]	**16,186 (-) [0,000]**	2,004 (-) [0,157]	**2,880 (-) [0,090]**	0,340 (-) [0,560]	**2,862 (-) [0,091]**	**9,190 (-) [0,002]**
h=4	**7,848 (-) [0,005]**	**3,302 (-) [0,069]**	**18,639 (-) [0,000]**	**6,768 (-) [0,009]**	**6,941 (-) [0,008]**	**3,457 (-) [0,063]**	**23,815 (-) [0,000]**	**3,918 (-) [0,048]**	2,015 (-) [0,156]	0,412 (-) [0,521]	**3,857 (-) [0,050]**	**9,409 (-) [0,002]**
h=5	**10,247 (-) [0,001]**	2,474 (-) [0,116]	**19,944 (-) [0,000]**	**5,728 (-) [0,017]**	**5,482 (-) [0,019]**	**8,437 (-) [0,004]**	**24,951 (-) [0,000]**	**5,895 (-) [0,015]**	1,008 (-) [0,315]	0,168 (-) [0,682]	**4,320 (-) [0,038]**	**8,014 (-) [0,005]**
h=6	**11,574 (-) [0,001]**	2,389 (-) [0,122]	**21,302 (-) [0,000]**	**5,563 (-) [0,018]**	**4,816 (-) [0,028]**	**9,397 (-) [0,002]**	**23,450 (-) [0,000]**	**5,747 (-) [0,017]**	0,343 (-) [0,558]	0,011 (-) [0,916]	**4,384 (-) [0,036]**	**4,426 (-) [0,035]**
h=7	**11,846 (-) [0,001]**	**3,030 (-) [0,082]**	**23,141 (-) [0,000]**	**6,406 (-) [0,011]**	**6,085 (-) [0,014]**	**12,823 (-) [0,000]**	**22,322 (-) [0,000]**	**5,316 (-) [0,021]**	0,054 (-) [0,816]	0,014 (-) [0,906]	**4,365 (-) [0,037]**	**4,399 (-) [0,036]**
h=8	**12,794 (-) [0,000]**	**3,860 (-) [0,049]**	**27,423 (-) [0,000]**	**8,011 (-) [0,005]**	**6,647 (-) [0,010]**	**18,473 (-) [0,000]**	**23,806 (-) [0,000]**	**4,941 (-) [0,026]**	0,253 (+) [0,615]	0,004 (-) [0,950]	**4,659 (-) [0,031]**	**6,883 (-) [0,009]**
h=9	**12,246 (-) [0,000]**	**4,052 (-) [0,044]**	**30,953 (-) [0,000]**	**9,666 (-) [0,002]**	**8,612 (-) [0,003]**	**26,778 (-) [0,000]**	**24,507 (-) [0,000]**	**4,635 (-) [0,031]**	0,813 (+) [0,367]	0,015 (-) [0,902]	**4,092 (-) [0,043]**	**12,402 (-) [0,000]**
h=10	**11,131 (-) [0,001]**	**3,929 (-) [0,047]**	**34,808 (-) [0,000]**	**12,050 (-) [0,001]**	**7,231 (-) [0,007]**	**35,103 (-) [0,000]**	**27,193 (-) [0,000]**	**4,445 (-) [0,035]**	1,406 (+) [0,236]	0,024 (-) [0,876]	**3,240 (-) [0,072]**	**13,081 (-) [0,000]**
h=11	**11,414 (-) [0,001]**	**3,540 (-) [0,060]**	**38,325 (-) [0,000]**	**14,096 (-) [0,000]**	**6,065 (-) [0,014]**	**40,769 (-) [0,000]**	**29,899 (-) [0,000]**	**3,175 (-) [0,075]**	1,971 (+) [0,160]	0,036 (-) [0,849]	2,400 (-) [0,121]	**12,936 (-) [0,000]**
h=12	**8,822 (-) [0,003]**	**3,403 (-) [0,065]**	**39,136 (-) [0,000]**	**16,023 (-) [0,000]**	**4,567 (-) [0,033]**	**35,856 (-) [0,000]**	**31,063 (-) [0,000]**	2,031 (-) [0,154]	1,835 (+) [0,175]	0,050 (-) [0,823]	1,676 (-) [0,195]	**11,899 (-) [0,001]**

Note: The table presents the results of the Giacomini and White (2006) test of unconditional predictive accuracy calculated during the excessive disinflation period (2011:01-2016:12). Under the null $H_0 : e^2_{t+h,M_A} - e^2_{t+h,M_B} = 0$. M_A, M_B and e_{t+h} denote the CB2 model, the competing model and the forecast error for horizon h, respectively. In the table I report the test statistics, the sign (in parentheses) indicating, whether model M_A delivers more accurate (as denoted by '−') or less accurate (as denoted by '+') forecasts and the p-value corresponding to the test statistic (in square brackets). Statistically more accurate forecast from the model M_A and M_B are reported in bold and italics, respectively. Source: own calculations.

In this restricted out-of-sample evaluation period the ANN improves upon the AO benchmark throughout the forecast horizon. However, the gain in the forecast accuracy is statistically insignificant for $h \leq 11$ and only marginally insignificant for $h = 12$. Moreover, the negative bias of the ANN model remains considerably smaller than in the case of the AO model. Overall, once again the AO benchmark has proved to be too hard to beat.

An important feature of the forecasts from the ANN model is that their accuracy and bias is not affected by restricting the sample to a period of structural change. While the negative bias of the competing models rises considerably indicating that a systematic error in these forecasts is present during the studied period, the bias of the ANN model remains very low. Moreover, the accuracy of the forecasts as measured by the RMSFE for all horizons even increases. For the nowcast, the forecast errors are 10% lower, whereas throughout all the forecasting horizons the RMSFE statistics is on average 6% lower. It is safely to assume, that the thick modelling combined with the use of bootstrap aggregating make the whole artificial neural network model more immune to structural changes.

Lastly, given the fact that the combination forecast of several linear and non-linear frameworks delivers further gains in the forecasting accuracy I examine the whether this combination produces statistically significant small average loss. Hence, I apply the Giacomini and White (2006) test considering in the null hypothesis the CB2 model as M_A and a competing benchmark M_B. In particular, I am interested whether this combination supplies sufficiently better forecasts to beat the demanding AO benchmark. Table 6 and 7 present the results of this procedure in a respective manner to the table 3.

Ex-post I conclude that the combined model pooling the forecasts from the univariate and multivariate, linear and non-linear beats the AO benchmark in the majority of cases for both the whole out-of-sample evaluation period as well as the excessive disinflation period. Moreover, during the evaluation period 2012:01-2016:12 this combination is able to improve in a statistically significant manner over the considered univariate approaches for almost all distinguished forecast horizons and for several cases outperforms also multivariate approaches. The gains in the predictive accuracy are substantial.

I conclude that in order to obtain accurate inflation forecasts one should consider combining several univariate and multivariate approaches with possibly diverse underlying assumptions. Let me underline here that the models entering the CB2 specification have in fact very little in common. Firstly, the ANN model is not constrained by stationarity assumption and its purpose is to recognize potentially non-linear patterns between the explanatory variable and the dependent variable by estimating weights using a specific gradient-based learning algorithm. Secondly, the SZ model implies that the priors on the seasonal variables are loose letting the data decide whether seasonality plays an important role in determining the development in the headline inflation. Thirdly, the VI model imposes a very tight prior on the unconditional mean of the model which results in accounting for the slowly evolving local inflation mean. When pooled together, the forecast accuracy of this model combination is superior in the majority of cases.

4.3. Robustness check

In this section I present the results of the robustness analysis. I check whether changing one of the key parameters of the studied model influences to a substantial extent the accuracy of the forecasts evaluated on the whole out-of-sample period (2011:01-2016:12) and during the excessive disinflation (2012:01-2016:12). Table 8 presents the RMSFE and the MFE (in parentheses) of the different specifications of the model calculated for the whole evaluation period.

Several interesting conclusions arise. Firstly, no single specification delivers the most accurate forecasts for all horizons as measured by the RMSFE and almost all specifications present small, negative bias indicating that the forecast on average pointed to higher than actual inflation. Secondly, the examined alternative specifications improve only marginally upon the baseline benchmark delivering a gain in accuracy not larger than 4%. It is thus safely to assume that no alternative specification statistically outperforms the baseline model for any horizon. Thirdly, the performance of the model in the out-of-sample period is insensitive to the amount of the information distributed to the learning set (as restricted by the parameter δ_1) and the initial parameter of the resilient propagation algorithm (Δ_{ij}). However, including more common unobserved factors leads to a substantial deterioration in the forecast accuracy. The increase in the RMSFE for $h = 12$ amounts to around 72% in comparison to the benchmark specification of the model when five common factors are included. Interestingly, further increase in the common factors to the amount which explains at least 80% of the variability of each vintage $\mathbb{X}_v$ increases marginally the accuracy of the nowcast and reduces the bias of the forecast. Still, expect the forecast horizon $h = 1$, the increase in the RMSFE statistic is substantial indicating the inferior predictive ability. Fourthly, small average complexity of the hidden layer leads to satisfactory accuracy – increasing the parameter γ governing the expected value in the Poisson distribution to augment the fraction of more complex non-linear artificial neural networks improves only marginally the performance for several horizons (the reduction in the RMSFE is not larger than 4%). Moreover, it appears that the regularization component has been set a little too tight in the baseline specification – increasing θ to 3 delivers small increases in the accuracy for almost all horizons and does not influence the average bias of the model. Diminishing the number of the pseudo-new learning sets does not severely decreases the accuracy of the model (in the worst scenario the increase in the RMSFE statistic for $h = 12$ equals to around 10%). However, there is one vital exception.

A distinctively inferior forecast delivers the specification XIV which assumes that there is neither random distribution of observations between the learning set and the validation set nor bootstrap aggregating. This specification and the results from this model need to be addressed here in more detail. In this approach the first 70% of the observations in a given vintage form the learning set and the remaining 30% constitute the validation set (the time succession of the observations is not disturbed). In total 10 000 models on a single learning set are estimated which are then used to obtain forecasts.

This model's forecast is supremely worse both in terms of the accuracy as measured by the RMSFE as well as the bias. For horizon $h = 12$ the average root mean squared forecasting error is almost twice as high and the bias is positive and substantial, indicating that on average the actual inflation has been much higher than predicted by the model. Thorough inspection of all forecast paths leads to the conclusion that the model is heavily overfitted and lost its generalization properties. The results for this specification exemplify the necessity to include methods aimed at stabilizing unstable learning procedures.

Apart from analysing the accuracy measured by the combination forecast, I investigate the quasi density of the point forecasts from all individual models for each horizon. Based on the individual point forecasts from all bagged artificial neural networks I use a kernel density function to approximate the density and evaluate it using the conditional ranked probability score. Table 9 presents the Crps of the different specifications of the Ann.

The conclusions hold once the CRPS statistic is compared – the baseline specification delivers competitive CRPS measures, whereas including more common factors or resigning from bagging deteriorates severely the quality of the density forecast. The sensitivity of the model's results with respect to changing other key parameters is limited. Again, selecting a slightly higher θ that penalizes the weights less seem to marginally improve the quality of the forecasts.

Interestingly, in the disinflation period the CRPS measure increases for horizons $h \leq 4$ and decreases sometimes quite substantially for further horizons. It indicates that during the disinflation period the quality of the forecasts generated for longer horizons improved at the expense of marginally worse nowcasts. As a result, these models appear to be an interesting supplement to the forecasting toolbox in the times of unorthodox inflation developments.

Table 8: The root mean square forecasting errors and mean forecasting errors of the studies specifications of the model in the whole evaluation period

	Specification													
Horizon	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
h=1	0,286 (0,006)	1,006 (0,004)	1,003 (0,015)	1,018 (0,012)	0,980 (0,004)	0,982 (-0,001)	0,999 (0,001)	1,011 (0,013)	1,021 (-0,001)	0,998 (0,012)	1,026 (0,011)	0,997 (0,011)	0,993 (0,006)	1,150 (0,135)
h=2	0,422 (0,006)	1,008 (0,005)	1,009 (0,024)	1,090 (0,018)	1,008 (0,002)	1,003 (-0,011)	1,008 (-0,007)	1,002 (0,018)	1,023 (-0,009)	0,987 (0,015)	1,031 (0,016)	1,023 (0,017)	0,993 (0,005)	1,286 (0,265)
h=3	0,600 (0,007)	0,991 (0,006)	0,988 (0,031)	1,121 (0,020)	1,062 (0,004)	0,994 (-0,017)	0,979 (-0,017)	0,982 (0,022)	1,005 (-0,016)	0,966 (0,014)	1,008 (0,020)	0,996 (0,021)	0,985 (0,002)	1,308 (0,396)
h=4	0,720 (-0,011)	0,996 (-0,014)	0,984 (0,017)	1,186 (0,001)	1,122 (-0,016)	0,992 (-0,041)	0,971 (-0,046)	0,975 (0,009)	1,000 (-0,042)	0,970 (-0,006)	0,993 (0,005)	0,995 (0,004)	1,003 (-0,016)	1,379 (0,514)
h=5	0,847 (-0,028)	0,998 (-0,039)	0,991 (0,004)	1,268 (-0,016)	1,190 (-0,031)	1,002 (-0,069)	0,982 (-0,070)	0,988 (0,000)	0,996 (-0,069)	0,969 (-0,027)	1,014 (-0,008)	1,012 (-0,010)	1,015 (-0,035)	1,432 (0,638)
h=6	0,937 (-0,053)	1,015 (-0,065)	0,987 (-0,012)	1,344 (-0,041)	1,243 (-0,050)	1,003 (-0,101)	0,962 (-0,107)	0,987 (-0,016)	0,990 (-0,102)	0,964 (-0,052)	1,022 (-0,030)	1,006 (-0,033)	1,018 (-0,060)	1,501 (0,758)
h=7	1,006 (-0,069)	1,017 (-0,074)	0,985 (-0,013)	1,426 (-0,050)	1,305 (-0,049)	1,012 (-0,123)	0,962 (-0,130)	0,990 (-0,018)	0,987 (-0,126)	0,975 (-0,062)	1,034 (-0,037)	1,007 (-0,042)	1,024 (-0,072)	1,603 (0,897)
h=8	1,083 (-0,085)	1,026 (-0,089)	0,991 (-0,021)	1,476 (-0,070)	1,333 (-0,058)	1,021 (-0,147)	0,979 (-0,159)	1,005 (-0,024)	0,994 (-0,156)	0,988 (-0,073)	1,051 (-0,046)	1,010 (-0,054)	1,030 (-0,089)	1,682 (1,038)
h=9	1,143 (-0,105)	1,036 (-0,111)	1,003 (-0,032)	1,539 (-0,095)	1,357 (-0,064)	1,030 (-0,178)	0,990 (-0,195)	1,014 (-0,042)	1,001 (-0,192)	0,989 (-0,093)	1,066 (-0,063)	1,010 (-0,078)	1,039 (-0,115)	1,778 (1,179)
h=10	1,196 (-0,120)	1,039 (-0,129)	1,006 (-0,038)	1,590 (-0,119)	1,372 (-0,064)	1,036 (-0,208)	1,000 (-0,225)	1,020 (-0,050)	1,011 (-0,219)	0,993 (-0,104)	1,078 (-0,077)	1,010 (-0,089)	1,053 (-0,130)	1,878 (1,335)
h=11	1,259 (-0,157)	1,041 (-0,167)	1,016 (-0,060)	1,653 (-0,156)	1,371 (-0,096)	1,051 (-0,250)	1,012 (-0,268)	1,025 (-0,077)	1,021 (-0,265)	1,004 (-0,136)	1,096 (-0,102)	1,015 (-0,119)	1,063 (-0,165)	1,943 (1,478)
h=12	1,310 (-0,204)	1,035 (-0,221)	1,009 (-0,101)	1,721 (-0,209)	1,416 (-0,105)	1,041 (-0,310)	1,027 (-0,328)	1,020 (-0,115)	1,020 (-0,329)	0,993 (-0,184)	1,102 (-0,146)	1,006 (-0,167)	1,053 (-0,216)	1,999 (1,611)

Note: The table presents the root mean squared forecasting errors and the mean forecasting errors (in parentheses) for the examined specifications of the model during the whole out-of-sample evaluation period (2011:01-2016:12). For the baseline model the error statistics are reported in levels. For the other considered specifications the RMSFE is presented as a ratio of the alternative and the baseline model. A ratio greater than one indicates that the competing specification produces on average less accurate forecasts. Source: own calculations.

Table 9: The mean of the continuous ranked probability score of the studied specifications of the model in the whole evaluation period and during the excessive disinflation

	Specification													
Horizon	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
h=1	0,150 (1,036)	0,158 (1,026)	0,154 (1,010)	0,155 (1,045)	0,150 (0,999)	0,144 (1,020)	0,158 (1,019)	0,160 (1,009)	0,152 (1,039)	0,157 (1,019)	0,165 (1,014)	0,153 (1,043)	0,143 (1,019)	0,209 (0,902)
h=2	0,244 (1,033)	0,249 (1,051)	0,242 (1,067)	0,239 (1,114)	0,232 (1,040)	0,242 (1,060)	0,247 (1,071)	0,245 (1,051)	0,245 (1,083)	0,241 (1,082)	0,250 (1,070)	0,261 (1,036)	0,243 (1,075)	0,344 (0,933)
h=3	0,366 (1,025)	0,366 (1,039)	0,376 (1,018)	0,409 (0,954)	0,391 (0,968)	0,376 (1,009)	0,364 (1,038)	0,372 (1,002)	0,378 (1,071)	0,356 (1,039)	0,378 (1,033)	0,367 (1,041)	0,362 (1,023)	0,490 (0,922)
h=4	0,447 (1,035)	0,494 (0,983)	0,477 (0,995)	0,563 (0,933)	0,538 (0,865)	0,482 (1,003)	0,483 (0,978)	0,476 (0,964)	0,473 (1,063)	0,461 (1,011)	0,484 (0,994)	0,486 (0,997)	0,462 (1,031)	0,611 (0,911)
h=5	0,599 (0,909)	0,593 (0,938)	0,578 (0,974)	0,739 (0,906)	0,682 (0,845)	0,600 (0,963)	0,578 (0,976)	0,592 (0,932)	0,570 (1,022)	0,566 (0,969)	0,619 (0,929)	0,588 (0,974)	0,589 (0,983)	0,728 (0,903)
h=6	0,662 (0,923)	0,666 (0,940)	0,665 (0,934)	0,863 (0,920)	0,772 (0,850)	0,680 (0,945)	0,641 (0,971)	0,665 (0,934)	0,654 (0,975)	0,637 (0,958)	0,683 (0,926)	0,669 (0,939)	0,676 (0,960)	0,832 (0,902)
h=7	0,696 (0,938)	0,700 (0,946)	0,700 (0,947)	0,971 (0,920)	0,853 (0,856)	0,717 (0,963)	0,678 (0,982)	0,692 (0,946)	0,692 (0,971)	0,668 (0,973)	0,728 (0,933)	0,701 (0,947)	0,722 (0,960)	0,934 (0,895)
h=8	0,738 (0,942)	0,748 (0,949)	0,744 (0,940)	1,082 (0,930)	0,941 (0,857)	0,762 (0,967)	0,719 (0,993)	0,734 (0,951)	0,720 (0,984)	0,711 (0,977)	0,775 (0,932)	0,741 (0,950)	0,766 (0,964)	1,053 (0,891)
h=9	0,752 (0,966)	0,771 (0,964)	0,768 (0,965)	1,187 (0,937)	0,995 (0,864)	0,785 (0,993)	0,741 (1,017)	0,759 (0,967)	0,752 (0,987)	0,727 (0,999)	0,807 (0,943)	0,759 (0,971)	0,790 (0,984)	1,165 (0,888)
h=10	0,764 (0,957)	0,789 (0,953)	0,786 (0,956)	1,274 (0,948)	1,013 (0,844)	0,802 (0,985)	0,757 (1,012)	0,782 (0,957)	0,778 (0,986)	0,744 (0,995)	0,837 (0,929)	0,771 (0,967)	0,814 (0,980)	1,281 (0,883)
h=11	0,770 (0,952)	0,810 (0,943)	0,816 (0,960)	1,384 (0,955)	1,045 (0,829)	0,829 (0,988)	0,773 (1,012)	0,805 (0,956)	0,810 (0,989)	0,760 (0,991)	0,876 (0,924)	0,791 (0,968)	0,843 (0,981)	1,381 (0,880)
h=12	0,794 (0,950)	0,835 (0,940)	0,843 (0,961)	1,505 (0,957)	1,127 (0,838)	0,848 (0,987)	0,808 (1,016)	0,830 (0,957)	0,846 (0,989)	0,778 (0,994)	0,916 (0,927)	0,821 (0,973)	0,867 (0,979)	1,481 (0,883)

Note: The table presents the average of the continuous ranked probability score for all considered specifications of the model calculated for the whole out-of-sample evaluation sample (2011:01-2016:12) and the excessive disinflation period (2012:01-2016:12, in parentheses, presented as a ratio of the CRPS calculated for excessive disinflation and the whole out-of-sample period). Source: own calculations.

Finally, I inspect whether the method of combining the point forecasts from individual artificial neural networks influences the accuracy of the combination forecast. Table 10 presents the RMSFE and the MFE statistics of the baseline model obtained using five examined forecast combination techniques. It

Table 10: The root mean squared forecasting errors and the mean forecasting errors of different combination methods

	RMSE				
Horizon	mean	median	trimmed mean	weighted (1)	weighted (2)
h=1	0,2862 (0,006)	0,9995 (0,005)	0,9998 (0,006)	**0,9989 (0,006)**	1,0002 (0,006)
h=2	0,4222 (0,006)	0,9993 (0,005)	0,9997 (0,006)	**0,9984 (0,006)**	0,9998 (0,007)
h=3	0,6002 (0,007)	0,9988 (0,005)	0,9995 (0,006)	**0,9978 (0,006)**	0,9998 (0,007)
h=4	0,7196 (-0,011)	0,9987 (-0,014)	0,9996 (-0,012)	**0,9976 (-0,012)**	1,0003 (-0,011)
h=5	0,8471 (-0,028)	**0,9976 (-0,031)**	0,9995 (-0,029)	0,9980 (-0,029)	1,0003 (-0,027)
h=6	0,9365 (-0,053)	**0,9975 (-0,057)**	0,9997 (-0,055)	0,9981 (-0,055)	1,0000 (-0,053)
h=7	1,0064 (-0,069)	**0,9971 (-0,074)**	0,9997 (-0,071)	0,9980 (-0,071)	0,9995 (-0,069)
h=8	1,0832 (-0,085)	**0,9973 (-0,090)**	0,9998 (-0,087)	0,9980 (-0,087)	0,9997 (-0,085)
h=9	1,1428 (-0,105)	**0,9978 (-0,111)**	1,0001 (-0,107)	0,9981 (-0,107)	0,9996 (-0,104)
h=10	1,1962 (-0,120)	0,9981 (-0,127)	1,0004 (-0,123)	**0,9980 (-0,122)**	0,9998 (-0,119)
h=11	1,2586 (-0,157)	0,9991 (-0,165)	1,0008 (-0,160)	**0,9982 (-0,159)**	0,9997 (-0,156)
h=12	1,3100 (-0,204)	1,0004 (-0,213)	1,0010 (-0,208)	**0,9978 (-0,207)**	1,0003 (-0,203)

Note: The table presents the root mean squared forecasting errors and the mean forecasting errors (in parentheses) for the studied combinations of individual forecasts during the whole out-of-sample evaluation period. For the baseline model the error statistics are reported in levels. For the other considered specifications the RMSFE is presented as a ratio of the alternative specification and the baseline model. A ratio greater than one indicates that the competing combination scheme produces on average less accurate forecasts. For each horizon the lowest RMSFE error is reported in bold. Source: own calculations.

appears that the model's accuracy and bias is completely insensitive to the method of averaging point forecasts from the individual artificial neural networks. The differences in the RMSFE as reported by the ratio of the RMSFE from the model with equal weights to a selected alternative are negligible.

5. Conclusion and discussion

In this paper I have aimed at evaluating the accuracy of the inflation forecasts generated by a combination of thousands of bagged single hidden-layer feed-forward artificial neural network. I have proposed an approach heavily dependent on strictly non-linear models and methods commonly used in the machine learning framework. I have contrasted the quality of the inflation forecasts from the proposed model with a number of econometric approaches commonly employed for inflation forecasting purposes. The forecasting experiment has been conducted using the large dataset of pseudo real-time data and evaluated on the out-of-sample period of systematically falling and persistently low inflation.

Choosing the optimal architecture of the artificial neural network is normally a subject of tedious, time-consuming experiments and subjectivity which may lead to dubious conclusion. The motivation of this paper was to restrict the number of subjective assumptions to the absolute minimum and establish the forecasting accuracy of the artificial neural networks by presenting an agnostic approach. To get around the caveat of introducing substantial subjectivity, I utilize the thick modelling framework and estimate a large number of simple hidden-layer feed forward artificial neural networks with the number of the neurons in the hidden layer following a zero-truncated Poisson distribution. In order to prevent the model from severe overfitting I have utilized bootstrap aggregating and considered a loss function with an additional weight regularisation component in the learning algorithm. The point forecast of the model is pooled from 10 000 generated specifications.

Several interesting conclusions can be adduced from the paper. First, a forecast combination of bagged single hidden-layer artificial neural networks outperforms a number of competing models in the established forecasting horse race. Importantly, for a considerable amount of cases the improvement upon the competing benchmark is statistically significant. During the period of excessive inflation in Poland their predictive accuracy has been affected and their bias remained stable which proves them a suitable ingredient in the forecasting toolbox in times of considerable structural change. However, they have failed to improve in a statistically significant manner upon the difficult benchmark proposed by Atkeson and Ohanian (2001).

These results corroborate a universal consensus that a single model tailored to all problems does not exist. However, the results indicate also that combining several linear and non-linear, univariate and multivariate approaches differing with respect to the underlying assumptions, included information and the relative performance across the forecasting horizon can prove most beneficial for the forecasting accuracy. In the paper I have combined the linear BVAR with loose prior imposed on the seasonal dummies, the non-linear BVAR with a tight prior imposed on the unconditional mean and the ANN not restricted by the stationarity constraint. During both analysed periods this combination has outperformed in a statistically significant manner a panel of competing benchmarks, including the Atkeson and Ohanian (2001) benchmark.

Two final comments are in order. First, it has to be underlined that the proposed framework does not take time explicitly into account. A common machine learning technique imposes that during the preprocessing phase the data are distributed between the learning and the validation set based on the random sampling (Zhang et al., 1998). Moreover, once bootstrap aggregating is employed, the set of constructed tuples of the dependent and the explanatory variables is perturbed further. Obviously, certain inference is lost. While this may raise criticism due to the fact that certain temporal relations are disregarded in the modelling phase, I have outlined in the paper that resigning from these procedures results in the serious overfitting of the combined model and severely compromises the accuracy of the point forecast. Taking into consideration the relationships between observations in time has remained beyond the scope of this paper and is a topic for future research.

Secondly, in contrast to the standard approaches, the interpretability of the combination of a large number of highly non-linear individual models is non-existent and curse of the black-box prevails. It remains a considerable hurdle for policy makers interested in understanding the reasons behind certain price developments. Therefore, I argue that although such models should by no means play the first fiddle, including them in the forecasting toolbox should prove useful and lead to the increase in the forecast accuracy of the combination of utilized approaches, especially in the periods of structural changes.

References

Ahmed, N. K., A. F. Atiya, N. E. Gayar, and H. El-Shishiny (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews 29*(5-6), 594–621.

Andrawis, R. R., A. F. Atiya, and H. El-Shishiny (2011). Forecast combinations of computational intelligence and linear models for the {NN5} time series forecasting competition. *International Journal of Forecasting 27*(3), 672 – 688. Special Section 1: Forecasting with Artificial Neural Networks and Computational IntelligenceSpecial Section 2: Tourism Forecasting.

Ang, A., G. Bekaert, and M. Wei (2007). Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics 54*(4), 1163 – 1212.

Atkeson, A. and L. E. Ohanian (2001). Are Phillips curves useful for forecasting inflation? *Quarterly Review* (Win), 2–11.

Baranowski, P., A. Leszczyńska, and G. Szafrański (2010). Krótkookresowe prognozowanie inflacji z użyciem modeli czynnikowych [short-term inflation forecasting using factor models]. *Bank i Kredyt 41*(4), 23–44.

Berg, T. O. and S. R. Henzel (2015). Point and density forecasts for the euro area using Bayesian VARs. *International Journal of Forecasting 31*(4), 1067–1095.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (favar) approach*. *The Quarterly Journal of Economics 120*(1), 387.

Binner, J. M., R. K. Bissoondeeal, T. Elger, A. M. Gazely, and A. W. Mullineux (2005). A comparison of linear forecasting models and neural networks: an application to euro inflation and euro divisia. *Applied Economics 37*(6), 665–680.

Binner, J. M., C. T. Elger, B. Nilsson, and J. A. Tepper (2006). Predictable non-linearities in u.s. inflation. *Economics Letters 93*(3), 323 – 328.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. New York, NY, USA: Oxford University Press, Inc.

Blinder, A. S., M. Ehrmann, M. Fratzscher, J. d. Haan, and D.-J. Jansen (2008). Central bank communication and monetary policy: A survey of theory and evidence. *Journal of Economic Literature 46*(4), 910–945.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*(2), 123–140.

Chen, X., J. Racine, and N. R. Swanson (2001). Semiparametric arx neural-network models with an application to forecasting inflation. *IEEE Transactions on neural networks 12*(4), 674–683.

Constâncio, V. (2015, August). Understanding Inflation Dynamics and Monetary Policy. Panel remarks at the Jackson Hole Economic Policy Symposium, Federal Reserve Bank of Kansas City.

Dagum, E. B. and P. A. Cholette (2006). *Benchmarking, temporal distribution, and reconciliation methods for time series*, Volume 186. Springer Science & Business Media.

Doz, C., D. Giannone, and L. Reichlin (2011). A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics 164*(1), 188 – 205. Annals Issue on Forecasting.

Duarte, C. and A. Rua (2007). Forecasting inflation through a bottom-up approach: How bottom is bottom? *Economic Modelling 24*(6), 941 – 953.

Edge, R. M. and R. S. Gurkaynak (2010). How Useful Are Estimated DSGE Model Forecasts for Central Bankers? *Brookings Papers on Economic Activity 41*(2 (Fall)), 209–259.

Faust, J. and J. H. Wright (2013). Chapter 1 - forecasting inflation. In G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 2, Part A of *Handbook of Economic Forecasting*, pp. 2 – 56. Elsevier.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2003). Do financial variables help forecasting inflation and real activity in the euro area? *Journal of Monetary Economics 50*(6), 1243 – 1255.

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

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

Goffe, W. L., G. D. Ferrier, and J. Rogers (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics 60*(1–2), 65 – 99.

Granger, C. W. and Y. Jeon (2004). Thick modeling. *Economic Modelling 21*(2), 323 – 343.

Groen, J. J. J., R. Paap, and F. Ravazzolo (2013). Real-time inflation forecasting in a changing world. *Journal of Business & Economic Statistics 31*(1), 29–44.

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

Igel, C. and M. Hüsken (2003). Empirical evaluation of the improved rprop learning algorithms. *Neurocomputing 50*, 105–123.

Inoue, A. and L. Kilian (2008). How useful is bagging in forecasting economic time series? a case study of u.s. consumer price inflation. *Journal of the American Statistical Association 103*(482), 511–522.

Jochmann, M., G. Koop, and R. W. Strachan (2010). Bayesian forecasting using stochastic search variable selection in a {VAR} subject to breaks. *International Journal of Forecasting 26*(2), 326 – 347. Special Issue: Bayesian Forecasting in Economics.

Khwaja, A., M. Naeem, A. Anpalagan, A. Venetsanopoulos, and B. Venkatesh (2015). Improved short-term load forecasting using bagged neural networks. *Electric Power Systems Research 125*, 109 – 115.

Kock, A. B. and T. Teräsvirta (2013, Spring). Forecasting the Finnish Consumer Price Inflation Using Artificial Neural Network Models and Three Automated Model Selection Techniques. *Finnish Economic Papers 26*(1), 13–24.

McAdam, P. and P. McNelis (2005). Forecasting inflation with thick models and neural networks. *Economic Modelling 22*(5), 848 – 867.

Moshiri, S. and N. Cameron (2000). Neural network versus econometric models in forecasting inflation. *Journal of Forecasting 19*(3), 201–217.

Moshiri, S., N. E. Cameron, and D. Scuse (1999). Static, dynamic, and hybrid neural networks in forecasting inflation. *Computational Economics 14*(3), 219–235.

Nakamura, E. (2005). Inflation forecasting using a neural network. *Economics Letters 86*(3), 373 – 378.

Nelson, M., T. Hill, W. Remus, and M. O'Connor (1999). Time series forecasting using neural networks: should the data be deseasonalized first? *Journal of Forecasting 18*(5), 359–367.

Orphanides, A. and V. W. Wieland (2008). Economic projections and rules of thumb for monetary policy. *Review* (Jul), 307–324.

Prieto, A., B. Prieto, E. M. Ortigosa, E. Ros, F. Pelayo, J. Ortega, and I. Rojas (2016). Neural networks: An overview of early research, current frameworks and new challenges. *Neurocomputing 214*, 242 – 268.

Rapach, D. E. and J. K. Strauss (2010). Bagging or combining (or both)? an analysis based on forecasting u.s. employment growth. *Econometric Reviews 29*(5-6), 511–533.

Refenes, A. N., A. Zapranis, and G. Francis (1994). Stock performance modeling using neural networks: A comparative study with regression models. *Neural Networks 7*(2), 375 – 388.

Riedmiller, M. and H. Braun (1993). A direct adaptive method for faster backpropagation learning: the rprop algorithm. In *IEEE International Conference on Neural Networks*, pp. 586–591 vol.1.

Rumler, F. and M. T. Valderrama (2010). Comparing the new keynesian phillips curve with time series models to forecast inflation. *The North American Journal of Economics and Finance 21*(2), 126 – 144. Special Issue: 50 Years of the Phillips Curve.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks 61*, 85 – 117.

Stella, A. and J. H. Stock (2012). A state-dependent model for inflation forecasting. International Finance Discussion Papers 1062, Board of Governors of the Federal Reserve System (U.S.).

Stelmasiak, D. and G. Szafrański (2016a). Forecasting the polish inflation using bayesian var models with seasonality. *Central European Journal of Economic Modelling and Econometrics 8*(1), 21–42.

Stelmasiak, D. and G. Szafrański (2016b). Inflation forecasting and information flow. does information flow matter to inflation? the importance of being timely on inflation. Technical report, Narodowy Bank Polski.

Stock, J. H. and M. W. Watson (1999). Forecasting inflation. *Journal of Monetary Economics 44*(2), 293 – 335.

Stock, J. H. and M. W. Watson (2003, September). Forecasting Output and Inflation: The Role of Asset Prices. *Journal of Economic Literature 41*(3), 788–829.

Stock, J. H. and M. W. Watson (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting 23*(6), 405–430.

Stock, J. H. and M. W. Watson (2007). Why has u.s. inflation become harder to forecast? *Journal of Money, Credit and Banking 39*, 3–33.

Swanson, N. R. and H. White (1997). Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. *International Journal of Forecasting 13*(4), 439 – 461.

Szafranek, K. (2017). Flattening of the new keynesian phillips curve: evidence for an emerging, small open economy. *Economic Modelling forthcoming*.

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

Timmermann, A. (2006). Chapter 4 forecast combinations. Volume 1 of *Handbook of Economic Forecasting*, pp. 135 – 196. Elsevier.

Zhang, G., B. E. Patuwo, and M. Y. Hu (1998). Forecasting with artificial neural networks:: The state of the art. *International Journal of Forecasting 14*(1), 35 – 62.

www.nbp.pl