Time Series Forecasting with Trend and Seasonal Patterns using NARX Network Ensembles

Hermansah^{1,2}, Dedi Rosadi^{1,*}, Abdurakhman¹, Herni Utami¹

¹Mathematics Study Program, Gadjah Mada University, Yogyakarta, 55281, Daerah Istimewa Yogyakarta, Indonesia

²Mathematics Education Study Program, Riau Kepulauan University, Batam, 29424, Kepulauan Riau, Indonesia

Received April 4, 2021; Revised June 1, 2021; Accepted June 15, 2021

Cite This Paper in the following Citation Styles

(a): [1] Hermansah, Dedi Rosadi, Abdurakhman, Herni Utami, "Time Series Forecasting with Trend and Seasonal Patterns using NARX Network Ensembles," *Mathematics and Statistics*, Vol.9, No.4, pp. 511-520, 2021. DOI: 10.13189/ms.2021.090410

(b): Hermansah, Dedi Rosadi, Abdurakhman, Herni Utami, (2021). *Time Series Forecasting with Trend and Seasonal Patterns using NARX Network Ensembles*. *Mathematics and Statistics*, 9(4), 511-520. DOI: 10.13189/ms.2021.090410

Copyright ©2021 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract In this research, we propose a Nonlinear Auto-Regressive network with exogenous inputs (NARX) model with a different approach, namely the determination of the main input variables using a stepwise regression and exogenous input using a deterministic seasonal dummy. There are two approaches in making a deterministic seasonal dummy, namely the binary and the sine-cosine dummy variables. Approximately half the number of input variables plus one is contained in the neurons of the hidden layer. Furthermore, the resilient backpropagation learning algorithm and the tangent hyperbolic activation function were used to train each network. Three ensemble operators are used, namely mean, median, and mode, to solve the overfitting problem and the single NARX model's weakness. Furthermore, we provide an empirical study using actual data, where forecasting accuracy is determined by Mean Absolute Percent Error (MAPE). The empirical study results show that the NARX model with binary dummy exogenous is the most accurate for trend and seasonal with multiplicative properties data patterns. For trend and seasonal with additive properties data patterns, the NARX model with sine-cosine dummy exogenous is more accurate, except the fact that the NARX model uses the mean ensemble operator. Besides, for trend and non-seasonal data patterns, the most accurate NARX model is obtained using the mean ensemble operator. This research also shows that the median and mode ensemble operators, which are rarely used, are more accurate than the mean ensemble operator for data that have trend and seasonal patterns. The median ensemble operator requires the least average computation time, followed by the mode ensemble operator. On the other hand, all of our proposed NARX models' accuracy consistently outperforms the exponential smoothing method and the ARIMA method.

Keywords Time Series, Forecasting, NARX Network, Stepwise Regression, Deterministic Seasonal Dummy, Ensemble Mean, Ensemble Median, Ensemble Mode

1 Introduction

Time series data forecasting is mainly done using statistical methods, namely Auto-Regressive Integrated Moving Average (ARIMA) or Box-Jenkins, Holt exponential smoothing, simple exponential smoothing, multiplicative Holt-Winter exponential smoothing, additive Holt-Winter exponential smoothing, etc. These method choices depend on various influencing aspects, namely the aspect of time, data patterns, the type of system model being observed, the desired level of forecasting accuracy, and others. However, these statistical methods are limited to linear models only [1, 2, 3]. Meanwhile, in many cases, forecasting time series data has a nonlinear tendency. Descriptions and applications for the nonlinearity test can be found at [4, 5].

Recently, a more flexible approach has been developed for modelling linear and nonlinear relationships, namely the Neural Networks (NN) model. The NN model is a widely used alternative because it does not require assumptions on the data that are often difficult to fulfil in the linear model [6, 7]. The most popular NN model for modelling and forecasting of time series data is the Nonlinear Auto-Regressive Neural Network (NARNN) model, also known as the Feed-Forward Neural Network (FFNN) or Back-Propagation Neural Network (BPNN) or Multi-Layer Perceptron (MLP) [8, 9]. On the other hand, the NN model is an unstable technique. It is because a small

change can result in a significant change in forecasting output. Many experiments have shown that the generalized results of a single NN model are not unique, meaning that a single NN model's solution is unstable. One alternative solution is found by assuming that the discarded NN model with inaccurate performance is considered to have potential as a candidate for the new NN model. Combining several NN models, is one of the effective ways of reducing the overfitting problem and the weaknesses of a single NN model in choosing a model. This combination of NN models is known as the ensemble NN [10, 11].

In this research, we propose an extension of the previous results of [11] on four sides: First, the forecasting model is considered for a more general class, namely the Nonlinear Auto-Regressive network with exogenous inputs (NARX) model, which previously applied the NARNN model. Second, the NARX model focuses on monthly time series data comprising of seasonal and trends patterns. A 12-period centered on moving average with time series data was employed on Cox-Stuart to identify the presence of trend. The test was carried out on the centered moving average to smoothen irregularities seasonally. Furthermore, the de-trended time series data was used to calculate and identify the seasonal indices, with a Friedman test used to determine its significant deviations. Seasonality or trend is identified in the time series data, and eliminated in the first differencing, in accordance with the research carried out by [12]. The time series data are then scaled linearly between -0.8 and 0.8 to facilitate the NARX model training. Besides, seasonal information is captured using a deterministic seasonal dummy as an exogenous input variables. Determination of the main input variables is done in the same way as [11] through the stepwise regression approach. Here, there are two approaches to make a deterministic seasonal dummy, namely the binary and the sine-cosine dummy variables.

Third, the NARX model with one hidden layer and the number of neurons in the hidden layer is half the number of input variables plus one. It is known that this similar approach is applied by [13] for the NARNN model. The tangent hyperbolic activation function and resilient back-propagation learning algorithm were used to train the NARX model. Fourth, the NARX model uses three ensemble operators, namely mean, median, and mode, to overcome the over-fitting problems and weaknesses of the single NARX model. Different ensemble sizes were evaluated from 10 networks to 100 networks. The results for a single NARX model by choosing the best, were not provided in this research, as there has been ample evidence in the literature that the ensemble is superior, see e.g. [14, 15].

This paper is structured as follows: This section explains our paper's background and contribution. In section 2, we provide a brief description of some theories required for a complete description of the paper. Section 3 provides an empirical study with actual data, namely the number of international airline passengers, the number of accidental deaths in the United States of America (USA), and Indonesian inflation rates. The measurement accuracy of forecasting is done by using the Mean Absolute Percent Error (MAPE) value. Besides, comparisons of MAPE values were carried out with ARIMA and exponential smoothing methods. Furthermore, the descrip-

tions of each were described in the studies [16] to see the increase in forecasting the NARX model using an ensemble.

2 Research Methods

2.1 Neural Networks

Neural Networks (NNs) was successfully enforced in forecasting univariate and multivariate time series. The Multi-Layer Perceptron (MLP), is the most utilized architecture with a linear and nonlinear functional relationship without preliminary assumptions on the underlying data generation process. Generally, MLPs are used to design Nonlinear Auto-Regressive NAR(p) and NARX(p)-processes. These models are used to forecast feed-forward architectures, while the external variables are used to code exogenous events. Furthermore, with a time series y , at a point in time t , a one-step ahead forecast $\hat{y}_{t+1}$ is determined using $p = I$ observations $y_t, y_{t-1}, \dots, y_{t-I+1}$ from I preceding points in time $t, t-1, t-2, \dots, t-I+1$, with I representing the number of input units of the MLP. The functional forms is as follows.

$$f(Y, w) = \beta_0 + \sum_{h=1}^H \beta_h g(\gamma_{0i} + \sum_{i=0}^I \gamma_{hi} y_i) \quad (1)$$

where $Y = [y_i, \dots, y_{t-I+1}]$ is the vector of the observed input time series. In addition, the network weights are $\gamma = [\gamma_{11}, \gamma_{12}, \dots, \gamma_{HI}]$ and $w = (\beta, \gamma)$, $\beta = [\beta_1, \beta_2, \dots, \beta_h]$ for the hidden input and output layers, respectively. The β_0 and γ_{0i} are the biases of each neuron, while the I and H are the number of input and hidden units in the network. The g is a non-linear transfer function, which is known as the logistic or hyperbolic tangent function [11].

2.2 NARX Model

The Nonlinear Auto-Regressive model with exogenous input (NARX) has been reported to be very essential to the discrete-time nonlinear systems and also defined using the following mathematical relationship [17]:

$$\begin{aligned} y(t+1) &= f(y(t), y(t-1), \dots, y(t-n_y+1); u(t), u(t-1), \\ &\quad \dots, u(t-n_u+1); \mathbf{w}) \\ &= f(\mathbf{y}(t); \mathbf{u}(t); \mathbf{w}) \end{aligned} \quad (2)$$

where $u(t)$ and $y(t)$ indicate the model's input and output at time t , respectively, $n_u \geq 1$ and $n_y \geq 1$ ($n_y \geq n_u$) represent input and output-memory orders. $\mathbf{w}$ and f are the weights matrix and nonlinear function used to estimate the Multi-Layer Perceptron (MLP) [18].

Basically, the NARX network is prepared using one out of two models [19]. The first is the parallel architecture (or with feedback), where the output is the feedback for the feed-forward neural networks input, being part of the standard architecture:

$$\begin{aligned} \hat{y}(t+1) &= \hat{f}(\hat{y}(t), \hat{y}(t-1), \dots, \hat{y}(t-n_y+1); u(t), u(t-1), \\ &\quad \dots, u(t-n_u+1); \mathbf{w}) \\ &= \hat{f}(\mathbf{y}_p(t); \mathbf{u}(t); \mathbf{w}) \end{aligned} \quad (3)$$

The second model is the series-parallel architecture (or parallel architecture without feedback), where there is formation of the output's regressors only through the use of the output actual values:

$$\begin{aligned}\hat{y}(t+1) &= \hat{f}(y(t), y(t-1), \dots, y(t-n_y+1); u(t), u(t-1), \dots, u(t-n_u+1); \mathbf{w}) \\ &= \hat{f}(\mathbf{y}_{sp}(t); \mathbf{u}(t); \mathbf{w})\end{aligned}\quad (4)$$

As previously stated, the NARX network inputs are included in the system's inputs and outputs regressors while the time-series is used as the output channels without the measured input. This model's ability to forecast is, therefore, limited in its application for time-series data without the input regressors, due to the elimination of the tapped-delay line over the input signal and this would further lead to the reduction of NARX to plain focused time-delay neural network architecture [20, 21]:

$$\hat{y}(t+1) = f(y(t), y(t-1), \dots, y(t-n_y+1); \mathbf{w}) \quad (5)$$

According to [22] and [23], a simple strategy was proposed in line with the Takens Embedding Theorem as a solution to the problem by providing the opportunity for the full exploitation of the original NARX network computational abilities in predicting nonlinear time-series and this, firstly, include defining the input signal regressors, denoted by $u(t)$, through the delay-embedding coordinates:

$$u(t) = (y(t), y(t-\tau), \dots, y(t-(d_E-1)\tau)) \quad (6)$$

where $d_E = n_u$ is the embedding dimension and τ is embedding delay.

Secondly, due to the possibility of training the NARX network in two different architectures, the output signal regressors $y(t)$ are presented as shown in the following relationships:

$$y_{sp}(t) = (y(t), y(t-1), \dots, y(t-n_y+1)) \quad (7)$$

$$y_p(t) = (\hat{y}(t), \hat{y}(t-1), \dots, \hat{y}(t-n_y+1)) \quad (8)$$

where the output regressor $y(t)$ for the parallel architecture without feedback in (7) has n_y actual time series past values while those with feedback in (8) has n_y estimated ones.

These outputs are values of $y(t+1)$ that have been previously estimated for a network that has been trained effectively and are required to follow the prognostic relationships applied using NARX network. This is represented as follows:

$$\hat{y}(t+1) = f(\mathbf{y}_{sp}(t); \mathbf{u}(t); \mathbf{w}) \quad (9)$$

$$\hat{y}(t+1) = f(\mathbf{y}_p(t); \mathbf{u}(t); \mathbf{w}) \quad (10)$$

The NARX networks trained in line with equations (9) and (10) are, therefore, represented as NARX-SP and NARX-P networks, respectively. This research was, therefore, focused on NARX models with parallel architecture and without feedback.

2.3 Exogenous Dummy Variables

According to [24], creating additional features in form of dummy variables is one of the methods used to capture deterministic trend without selecting lagged inputs. The conventional approach used to model deterministic seasonal patterns are $S-1$ binary dummy variables with period of t , where S denotes the seasonal length. Meanwhile, for high frequency data, long input vectors used $S-1$ additional time series. A set of sine and cosine dummy variables, captures deterministic seasonal elements of the time series [25]. In addition, two inputs including $x_{s,1}$ and $x_{s,2}$ were used to encode seasonality with variables created using $\sin(t)$ and $\cos(t)$ for an explicit representation of the point within an identified seasonality of length s , (see e.g. [24, 25]) with

$$x_{s,1}(t) = \sin\left(\frac{2\pi t}{S}\right) \quad \text{and} \quad x_{s,2}(t) = \cos\left(\frac{2\pi t}{S}\right). \quad (11)$$

Furthermore, $x_{s,1}$ and $x_{s,2}$ are used to determine the sine-cosine-pairs for each s and the input vector for long and multiple seasonalities. This study applied the two conventional approaches as exogenous input variables in the NARX model.

2.4 NARX Modelling

NARX modelling generally consists of five steps.

- Step 1: Preprocessing of the time series started with seasonality, checking the time series trend, and applying seasonal difference or the difference if the time series contains seasonality or trend patterns. The time series was further scaled into the range of $[-0.8, 0.8]$.
- Step 2: The following strategy was used to obtain the autoregressive lags for the main time series and the exogenous variables. If the main time series frequency is equal to m , all lags from 1 to m were considered as possible lag numbers. For example, for monthly data, 1 to 12 was considered. The lag order of the exogenous variables was assumed to equal the main time series, and the significant lags were selected using the stepwise AIC method. Here, the exogenous variables include the seasonal dummy variables. There are two approaches to make seasonal dummy variables, namely the binary and the sine-cosine dummy variables.
- Step 3: The number of neurons in the hidden layer was chosen to be half of the total input in Step 2 plus 1.
- Step 4: The NARX networks obtained in Step 3 were fitted from 10 to 100 times using different random starting weights, and the forecasts obtained were combined using the ensemble approach (mean, median, and mode) to produce the final forecast.
- Step 5: The recursive or iterative strategy was used for multi-step ahead forecasts.

Further detail, which also shows the above algorithm's applicability for the NARNN model, is provided in [11].

2.5 Ensemble Operators

The application of $\hat{y}_{mt}$ as the forecast from model m for time t , where $m = 1, \dots, M$ and M represents the number of forecasts available for combination in an ensemble forecast $\tilde{y}_t$. There was a discussion on the formation of $\tilde{y}_t$ through the use of the mean, median, and mode operators in this section and a unimodal distribution was assumed to achieve dependable results.

2.5.1 Mean Ensemble

One of the prevalent ways to determine the central tendency of a set of data that is weighted or unweighted is by using the mean ensemble method. Let w_m denote the forecast's weight from model m . Therefore, using $0 \leq w_m \leq 1$ and $\sum_{m=1}^M w_m = 1$, the ensemble forecast t is calculated as follows.

$$\tilde{y}_t^{Mean} = \sum_{m=1}^M w_m y_{mt} \quad (12)$$

The combination produced for all $w_m = M^{-1}$ is unweighted with the attributes and limits of the mean known. It was observed to have some sensitivity to outliers and this makes it not to be reliable for skewed distributions. It is, however, possible to use winsorized or truncated means to obviate some of these problems by making the mean behave more closely to the median. Moreover, the maximum distance between these two measures is one standard deviation for distributions with finite variance and this has been reported to be true for sets of forecasts [26].

2.5.2 Median Ensemble

It is also possible for the median to either be unweighted or weighted but the weighted aspect is not usually applied. The calculation of median ensemble $\tilde{y}_t^{Median}$ is conducted by arranging $w_m y_{mt}$ and selecting a single value in the middle if M is not even or by determining the average of the two values in the middle when it is not odd. It is, however, important to note that it also feels the effect of non-symmetric distributions even though it is more robust than the mean.

2.5.3 Mode Ensemble

The mode ensemble is the most occurring value in a set of data and has been found not to be sensitive to outliers unlike the mean and the median. No formula was found to calculate unidentified continuous variables, but two methods are usually applied and these include data discretization and identification of the most frequent bin as well as kernel density estimation. The second approach was preferred in this research due to its effectiveness for continuous-valued nature of forecasts.

Kernel density estimation has been identified as a non-parametric method of calculating the probability density function of a random variable and, for the purpose of this research, the forecasts. It is possible to estimate the shape of the function using this method if some forecasts of distribution with

unknown density f are provided.

$$\hat{f}_{th}(x) = (Mh)^{-1} \sum_{m=1}^M K\left(\frac{x - \hat{y}_{mt}}{h}\right) \quad (13)$$

where $K(\cdot)$ is a function with $\int K(x)dx = 1$ known as the kernel, while $h \geq 0$ is the bandwidth. It is most times used as a unimodal symmetric density function and this makes the $\hat{f}_{th}(x)$ a density function itself and this is most times known as the Gaussian kernel $\phi(x)$:

$$\phi_h(x) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{x^2}{2h^2}} \quad (14)$$

Previous literature has proposed several alternative kernel functions but the one selected was observed to have a minimal effect on the results in most cases. Moreover, the kernel h bandwidth moderates the smoothing quantity and, at a higher value, more smoothing is provided. Therefore, the h selected is very important mainly due to the ability of under-smoothing or over-smoothing to give wrong density f estimation [27]. Meanwhile, the approximation by [28] is often used in practice

$$h = \left(\frac{4\hat{\sigma}^5}{3M}\right)^{\frac{1}{5}} \quad (15)$$

where $\hat{\sigma}$ represents the forecast samples' standard deviation and its estimate is most appropriate for Gaussian kernels. Moreover, [29] proposed a method for the automatic selection of bandwidth without being affected by the arbitrary standard reference rules used in previous studies. The preference of this method in calculating the mode ensemble is due to its ability to produce a bandwidth h allowing the ensemble's fast convergence and excellent performance.

The value x corresponding to the maximum value of density function estimates the mode for the true fundamental set of forecasts distribution and also serves as the mode ensemble value $\tilde{y}_{t+h}^{Mode}$. Meanwhile, a relatively numerous observations are needed for the appropriate display of the fundamental density by the kernel density estimation because using a small number usually leads to a bad approximation.

3 Result and Discussion

This research was conducted on three cases of actual data: the number of international airline passengers, the number of accidental deaths in the USA, and Indonesian inflation rates. The measurement accuracy of forecasting is done by using the Mean Absolute Percent Error (MAPE) value. MAPE is a measure of accuracy that is widely used in time series forecasting studies. A method performs very well when the MAPE value is below 10% and performs well when the MAPE value is between 10% and 20%. MAPE is defined as follows [30]:

$$MAPE = \frac{1}{N} \sum_{t=1}^N \left| \frac{A_t - F_t}{A_t} \right| \quad (16)$$

where N is the number of data, A_t is actual values at data time t , and F_t is forecast values at data time t .

3.1 Airline Passengers

Data on the number of international airline passengers, or better known as airline data, are popular data and are often used as an example of applying the time series method, among others in [1, 2, 6, 7]. The data used in this research are monthly airline passenger numbers from January 1949 to December 1960. Thus, the number of samples used is 144, and the last 12 data are used as the testing data. Monthly data of international airline passengers have an upward trend pattern and seasonal pattern with multiplicative properties, and graphically it can be seen in Figure 1.

This research aims to determine the monthly time forecasting series data using trend and seasonal patterns. The Cox-Stuart test was used to determine trends in a time series data on a 12-period moving average. The test was carried out on the moving average to smoothen effects due to irregularities. Furthermore, the de-trended time series data were used to calculate and identify seasonal indices, which were tested to determine their significant deviations through a Friedman test. The result showed that time series data identified through a trend is eliminated through the first differencing method. The time-series data are further scaled linearly between -0.8 and 0.8 to facilitate the NARX model's training.

Figure 1. Plot of real data, fitting in sample and out sample forecast using ARIMA, ETS, NARX 1 (the best model of NARX with binary dummy variables), and NARX 2 (the best model of NARX with sine-cosine dummy variables) for the airline passengers.

The NARX model's design is as follows: Determination of the main input variables is carried out in the same way as [11] through a stepwise regression approach, and exogenous input variables are made with a deterministic seasonal dummy. There are two approaches to make a deterministic seasonal dummy, namely the binary and the sine-cosine dummy variables. The number of neurons in the hidden layer is half the number of input variables plus one. Each network was qualified using the resilient backpropagation learning algorithm and the tangent hyperbolic activation function. Finally, the NARX model design uses three ensemble operators: mean, median, and mode, to solve the overfitting problem and the single NARX model's weaknesses. The ensemble size is evaluated from 10 networks to 100 networks.

Figure 2. Plot of real data and out sample forecast using ARIMA, ETS, NARX 1, and NARX 2 for the airline passengers.

Based on the NARX model's design above, the MAPE results for data on the number of international airline passengers can be seen in Tables 1 and 2. Overall, the results of NARX model with binary dummy variables in Table 1 are more accurate than the NARX model with sine-cosine dummy variables in Table 2. These differences in results indicate that the second approach as an exogenous dummy variable is less robust for seasonal data with multiplicative properties. However, all NARX models in this first case study consistently outperformed the exponential smoothing method and the ARIMA method. Exponential smoothing through the state-space approach obtained the best model using the Holt-Winter linear multiplicative method or ETS (M,A,M) with an AIC of 1395.16. Meanwhile, through the stepwise ARIMA approach, the best model is obtained using ARIMA (2,1,1)(0,1,0)[12] with an AIC of 1017.85. Graphic illustration in the form of plot of real data, fitting in sample and out sample forecast using NARX 1 (the best model of NARX with binary dummy variables), NARX 2 (the best model of NARX with sine-cosine dummy variables), ARIMA, and ETS can be seen in Figures 1 and 2.

Overall, the NARX model with mode ensemble operator (NARX-Mode) is an ideal model with the lowest MAPE value. The NARX-Mode performs better when the ensemble size is large enough. In particular, the ensemble size becomes very important for the accuracy of the mode ensemble operator. It is because the kernel density estimation becomes reliable when the number of observations is adequate, as discussed in Section 2. This research found that the rarely used median and mode operators are more accurate than the mean ensemble operator. The NARX-Median is slightly lower in performance but can perform well for small and large ensembles, meaning that the NARX-Median is relatively stable. The NARX-Median and NARX-Mean rank second and third with slight differences, respectively.

Table 1. MAPE results from the NARX model with binary dummy variables for the airline passengers

Ensemble size	NARX-Mean	NARX-Median	NARX-Mode	ARIMA	ETS
10	0.0079	0.0078	0.0082	0.0280	0.0285
20	0.0077	0.0076	0.0074	0.0280	0.0285
30	0.0079	0.0076	0.0085	0.0280	0.0285
40	0.0078	0.0075	0.0079	0.0280	0.0285
50	0.0076	0.0076	0.0068	0.0280	0.0285
60	0.0075	0.0073	0.0075	0.0280	0.0285
70	0.0073	0.0074	0.0067	0.0280	0.0285
80	0.0074	0.0075	0.0070	0.0280	0.0285
90	0.0077	0.0073	0.0070	0.0280	0.0285
100	0.0077	0.0075	0.0073	0.0280	0.0285

Table 2. MAPE results from the NARX model with sine-cosine dummy variables for the airline passengers

Ensemble size	NARX-Mean	NARX-Median	NARX-Mode	ARIMA	ETS
10	0.0101	0.0099	0.0104	0.0280	0.0285
20	0.0105	0.0093	0.0097	0.0280	0.0285
30	0.0102	0.0095	0.0100	0.0280	0.0285
40	0.0098	0.0098	0.0096	0.0280	0.0285
50	0.0097	0.0094	0.0100	0.0280	0.0285
60	0.0102	0.0099	0.0101	0.0280	0.0285
70	0.0101	0.0096	0.0096	0.0280	0.0285
80	0.0098	0.0100	0.0094	0.0280	0.0285
90	0.0104	0.0099	0.0092	0.0280	0.0285
100	0.0101	0.0099	0.0092	0.0280	0.0285

3.2 Accidental Deaths

Monthly data on accidental deaths in the United States of America (USA) can be seen graphically, as shown in Figure 3. Monthly data from January 1973 to December 1978 were used as the training data, and the data from January 1979 to June 1979 were used as testing data. The data used in this research is popular due to its complex pattern. Furthermore, research by Brockwell and Davis [3] was carried out by applying the ARIMA and exponential smoothing models.

Figure 3. Plot of real data, fitting in sample and out sample forecast using ARIMA, ETS, NARX 1, and NARX 2 for the accidental deaths in the USA.

The design is the same as in the first case study. The MAPE results on the NARX model for monthly data of accidental deaths in the USA can be seen in Tables 3 and 4. The results of NARX model with sine-cosine dummy variables (second approach) in Table 4 are, overall, more accurate for NARX-Median and NARX-Mode. However, the same thing is not

Figure 4. Plot of real data and out sample forecast using ARIMA, ETS, NARX 1, and NARX 2 for the accidental deaths in the USA.

true for the NARX-Mean. The NARX-Mean accuracy with the first approach is slightly better. These results also indicate that the second approach for NARX-Median and NARX-Mode is robust for seasonal data with additive properties. Overall, the NARX-Mode is the most accurate model with the lowest MAPE value. The NARX-Median ranks second, followed by the third NARX-Mean. The accuracy of all NARX models in the second case study consistently outperformed the exponential smoothing method and the ARIMA method. Exponential smoothing as the best model uses the Holt-Winter linear additive method or ETS (A,A,A) with an AIC of 1240.18. Meanwhile, the best ARIMA uses ARIMA (0,1,1)(0,1,1)[12] with an AIC of 942.27. Graphic illustration in the form of plot of real data, fitting in sample and out sample forecast using NARX 1 (the best model of NARX with binary dummy variables),

Table 3. MAPE results from the NARX model with binary dummy variables for the accidental deaths in the USA

Ensemble size	NARX-Mean	NARX-Median	NARX-Mode	ARIMA	ETS
10	0.0028	0.0030	0.0023	0.0239	0.0236
20	0.0029	0.0025	0.0026	0.0239	0.0236
30	0.0032	0.0026	0.0022	0.0239	0.0236
40	0.0031	0.0025	0.0023	0.0239	0.0236
50	0.0031	0.0025	0.0020	0.0239	0.0236
60	0.0029	0.0024	0.0019	0.0239	0.0236
70	0.0030	0.0024	0.0019	0.0239	0.0236
80	0.0030	0.0025	0.0020	0.0239	0.0236
90	0.0029	0.0024	0.0021	0.0239	0.0236
100	0.0029	0.0025	0.0020	0.0239	0.0236

Table 4. MAPE results from the NARX model with sine-cosine dummy variables for the accidental deaths in the USA

Ensemble size	NARX-Mean	NARX-Median	NARX-Mode	ARIMA	ETS
10	0.0034	0.0024	0.0022	0.0239	0.0236
20	0.0037	0.0023	0.0019	0.0239	0.0236
30	0.0035	0.0024	0.0017	0.0239	0.0236
40	0.0034	0.0024	0.0014	0.0239	0.0236
50	0.0034	0.0023	0.0016	0.0239	0.0236
60	0.0035	0.0020	0.0012	0.0239	0.0236
70	0.0030	0.0023	0.0013	0.0239	0.0236
80	0.0034	0.0020	0.0016	0.0239	0.0236
90	0.0034	0.0021	0.0014	0.0239	0.0236
100	0.0032	0.0021	0.0014	0.0239	0.0236

NARX 2 (the best model of NARX with sine-cosine dummy variables), ARIMA, and ETS can be seen in Figures 3 and 4.

3.3 Indonesian Inflation

The third case study is the data on the inflation rate in Indonesia. Inflation rate data in Indonesia is the monthly data that is observed from January 2007 to February 2018 or consists of 134 observations. Model training is carried out on the first 129 data (or training data), and the last five data is used as testing data. Graphically, data on the inflation rate in Indonesia can be seen in Figure 5. The data graph shows that there is a downward and upward trend over a certain period.

Figure 5. Plot of real data, fitting in sample and out sample forecast using ARIMA, ETS, and NARX for the Indonesian inflation with the exogenous variable being the interest rate of the Indonesian Central Bank.

The NARX model's design in the third case study is not the same as in the first and second case studies. It means that it

does not consider the deterministic seasonal dummy. Here, the exogenous variables of the NARX model use a second correlation approach such as [31] in modelling a single NARX. The exogenous variable with the second correlation chosen is the interest rate data of the Indonesian Central Bank. The lag order for the interest rate data was assumed to equal the inflation rate data. MAPE results from the NARX model for the third case study can be seen in Table 5. Overall, the third case study found that the NARX-Mean is an ideal model with the lowest MAPE value, followed by NARX-Median with little difference. Meanwhile, NARX-Mode ranks third, which previously found the most accurate in the first and second case studies. This indicates that NARX-Mode is less robust for non-seasonal data. However, all of the NARX models consistently outperformed the exponential smoothing method and the ARIMA method. Exponential smoothing using the Holt linear method or ETS (M,A,N) with an AIC of 648.19. While the ARIMA model uses ARIMA (0,1,1)(2,0,0)[12] with an AIC of 378.90. A graphic illustration in the form of plot of real data, fitting in sample and out sample forecast using ARIMA, ETS, and NARX, is shown in Figure 5.

On the other hand, the comparison of the computation time required for forecasting all ensemble sizes of each data as shown in Table 6. To show the comparison of different computation times, we summarize them into the average computation time. The average of computation time is shown in the form of seconds. The network is trained in parallel on a processor i5-8250U CPU with 1.8GHz. Table 6 shows the median operator requires the least average computation time compared to the mean or mode operators for each data. The mode operator ranks second with a slight difference. Meanwhile, the mean operator which is most often used requires the most average computation time.

Table 5. MAPE results from the NARX model for the Indonesian inflation

Ensemble size	NARX-Mean	NARX-Median	NARX-Mode	ARIMA	ETS
10	0.0465	0.0534	0.0508	0.0788	0.1047
20	0.0487	0.0508	0.0503	0.0788	0.1047
30	0.0489	0.0508	0.0530	0.0788	0.1047
40	0.0488	0.0497	0.0521	0.0788	0.1047
50	0.0491	0.0485	0.0513	0.0788	0.1047
60	0.0482	0.0484	0.0515	0.0788	0.1047
70	0.0497	0.0486	0.0492	0.0788	0.1047
80	0.0496	0.0488	0.0507	0.0788	0.1047
90	0.0483	0.0481	0.0517	0.0788	0.1047
100	0.0485	0.0486	0.0520	0.0788	0.1047

Table 6. Computational time comparison from the NARX models

Ensemble operator	Ensemble size	Airline passengers	Accidental deaths	Indonesian inflation
Mean	10	2.823 secs	1.346 secs	3.176 secs
Median		2.750 secs	1.117 secs	1.641 secs
Mode		2.631 secs	1.270 secs	1.739 secs
Mean	20	5.486 secs	2.347 secs	4.515 secs
Median		4.032 secs	2.364 secs	2.357 secs
Mode		5.286 secs	2.167 secs	3.209 secs
Mean	30	7.919 secs	3.491 secs	5.040 secs
Median		7.205 secs	3.236 secs	4.381 secs
Mode		7.426 secs	3.315 secs	4.787 secs
Mean	40	10.79 secs	5.156 secs	6.653 secs
Median		9.401 secs	5.115 secs	5.128 secs
Mode		9.602 secs	4.830 secs	6.176 secs
Mean	50	12.56 secs	6.064 secs	8.245 secs
Median		11.92 secs	6.312 secs	7.687 secs
Mode		11.99 secs	5.702 secs	6.932 secs
Mean	60	15.50 secs	7.592 secs	9.459 secs
Median		14.52 secs	7.809 secs	8.477 secs
Mode		15.24 secs	7.320 secs	9.177 secs
Mean	70	18.05 secs	8.929 secs	10.978 secs
Median		16.52 secs	8.675 secs	10.058 secs
Mode		17.11 secs	7.706 secs	10.836 secs
Mean	80	21.51 secs	9.772 secs	12.570 secs
Median		20.96 secs	9.930 secs	12.467 secs
Mode		19.81 secs	9.410 secs	11.477 secs
Mean	90	23.79 secs	10.576 secs	13.694 secs
Median		22.98 secs	10.084 secs	13.431 secs
Mode		22.04 secs	11.092 secs	13.047 secs
Mean	100	24.49 secs	11.933 secs	15.980 secs
Median		22.86 secs	11.555 secs	14.351 secs
Mode		23.46 secs	11.581 secs	15.737 secs

4 Conclusions

The first case study on the trend and seasonal patterns with multiplicative properties which shows all the NARX model approaches (NARX-Mean, NARX-Median, and NARX-Mode) with binary dummy are more accurate than the NARX model approaches with sine-cosine dummy. The second case study on the trend and seasonal patterns with additive properties obtained that the NARX models (NARX-Median and NARX-Modus) is more accurate with sine-cosine dummy. The same thing is not proper for NARX-Mean. Meanwhile, the third case study on the trend and non-seasonal patterns obtained the most accurate using NARX-Mean. Overall, NARX-Mode is the most accurate model with the lowest MAPE value for data that has both trend and seasonal patterns. However, the NARX-Mean is most accurate for data that has trend and non-seasonal patterns. The NARX-Mode performs better when the ensemble

size is large enough. The NARX-Median has slightly lower performance but it is stable enough for each data pattern and ensemble size. This case study also shows that the rarely used median and mode operators are more accurate than the mean operator. On the other hand, the median operator takes the least average computation time, followed by the mode operator in the second place with a slight difference. However, all NARX models in this study consistently outperformed the exponential smoothing method and the ARIMA method. Based on these results, we recommend the next researcher to use the dummy variables as exogenous input of the NARX model and use the median and mode operators in ensembles research and applications, which previously focused only on the mean operator. However, further research is needed to develop the NARX model of parallel architecture with feedback and weighted versions of the operators.

Acknowledgements

The authors are grateful to the Ministry of Education and Culture Republic of Indonesia for their financial support via Doctoral Research Grant (Penelitian Disertasi Doktor, PDD) 2021.

REFERENCES

- [1] G. E. P. Box, G. M. Jenkins. Time Series Analysis: Forecasting and Control, Holden Day, San Francisco, 1976.
- [2] R. J. Hyndman, A. B. Koehler, J. K. Ord, R. D. Snyder. Forecasting with Exponential Smoothing: The State Space Approach, Springer Science and Business Media, Germany, 2008.
- [3] P. J. Brockwell, R. A. Davis. Introduction to Time Series and Forecasting, Springer-Verlag, New York, 2002.
- [4] T. Teraesvirta, C. F. Lin, C. W. J. Granger. Power of the Neural Network Linearity Test, Journal of Time Series Analysis, Vol.14, No.2, 209-220, 1993.
- [5] T. H. Lee, H. White, C. W. J. Granger. Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests, Journal of Econometrics, Vol.56, No.3, 269-290, 1993.
- [6] J. Faraway, C. Chatfield. Time series forecasting with neural networks: A comparative study using the airline data, Journal of the Royal Statistical Society, Vol.47, No.2, 231-250, 1998.
- [7] Suhartono. Feedforward Neural Networks untuk Pemodelan Runtun Waktu, Universitas Gadjah Mada, Yogyakarta, 2007.
- [8] Hermansah, D. Rosadi, Abdurakhman, H. Utami. Comparison of ARIMA, neural network, and wavelet models for forecasting Indonesia sharia stock index, ISI World Statistics Congress, Kuala Lumpur, 286-293, 2019.
- [9] Hermansah, D. Rosadi, H. Utami, Abdurakhman, G. Darmawan. Hybrid MODWT-FFNN model for time series data forecasting, SEAMS Universitas Gadjah Mada, Yogyakarta, 1-8, 2019.
- [10] N. Kourentzes, D. K. Barrow, S. F. Crone. Neural network ensemble operators for time series forecasting, Expert Systems with Applications, Vol.41, No.9, 4235-4244, 2014.
- [11] Hermansah, D. Rosadi, Abdurakhman, H. Utami. Selection of input variables of nonlinear autoregressive neural network model for time series data forecasting, Media Statistika, Vol.13, No.2, 116-124, 2020.
- [12] G. P. Zhang, M. Qi. Neural network forecasting for seasonal and trend time series, European Journal of Operational Research, Vol.160, No.2, 501-514, 2005.
- [13] R. J. Hyndman, G. Athanasopoulos. Forecasting: Principles and Practice, OTexts, Monash University, 2018.
- [14] G. P. Zhang, V. L. Berardi. Time series forecasting with neural network ensembles: an application for exchange rate prediction, Journal of the Operational Research Society, Vol.52, No.1, 652-664, 2001.
- [15] D. K. Barrow, S. F. Crone, N. Kourentzes. An evaluation of neural network ensembles and model selection for time series prediction, International Joint Conference on Neural Networks, Barcelona, 1-8, 2010.
- [16] R. J. Hyndman, Y. Khandakar. Automatic time series forecasting: The forecast package for R, Journal of Statistical Software, Vol.27, No.3, 1-22, 2008.
- [17] H. T. Siegelmann, B. G. Horne, C. L. Giles. Computational capabilities of recurrent NARX neural networks, IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics, Vol.27, No.2, 208-215, 1997.
- [18] K. S. Narendra, K. Parthasarathy. Identification and control of dynamical systems using neural networks, IEEE Transactions on Neural Networks, Vol.1, No.1, 4-27, 1990.
- [19] C. Jiang, F. Song. Sunspot forecasting by using chaotic time-series analysis and NARX network, Journal of Computers, Vol.6, No.7, 1424-1429, 2011.
- [20] Y. Bengio, P. Simard, P. Frasconi. Learning long-term dependencies with gradient descent is difficult, IEEE Transactions on Neural Networks, Vol.5, No.2, 157-166, 1994.
- [21] T. Lin, B. G. Horne, P. Tino, C. L. Giles. Learning long-term dependencies in NARX recurrent neural networks, IEEE Transactions on Neural Networks, Vol.7, No.6, 1329-1338, 1996.
- [22] S. Ingrassia, I. Morlini. Advances in Data Analysis: Equivalent Number of Degrees of Freedom for Neural Networks, Springer Berlin Heidelberg, Berlin, 229-236, 2007.
- [23] E. Diaconescu. The use of NARX neural networks to predict chaotic time series, WSEAS Transactions on Computer Research, Vol.3, No.3, 182-191, 2008.
- [24] E. Ghysels, D. R. Osborn. The Econometric Analysis of Seasonal Time Series, Cambridge University Press, Cambridge, 2001.

- [25] S. F. Crone, N. Kourentzes. Feature selection for time series prediction: A combined filter and wrapper approach for neural networks, *Neurocomputing*, Vol.73, No.10, 1923–1936, 2010.
- [26] V. R. Jose, R. L. Winkler. Simple robust averages of forecasts: Some empirical results, *International Journal of Forecasting*, Vol.24, No.1, 163-169, 2008.
- [27] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.
- [28] B. W. Silverman. Using kernel density estimates to investigate multimodality, *Journal of the Royal Statistical Society*, Vol.43, No.1, 97-99, 1981.
- [29] Z. I. Botev, J. F. Grotowski, D. P. Kroese. Kernel density estimation via diffusion, *The Annals of Statistics*, Vol.38, No.5, 2916-2957, 2010.
- [30] R. J. Hyndman, A. B. Koehler. Another look at measures of forecast accuracy, *International Journal of Forecasting*, Vol.22, No.4, 679-688, 2006.
- [31] A. Sihabuddin, Subanar, D. Rosadi, E. Winarko. A second correlation method for multivariate exchange rates forecasting, *International Journal of Advanced Computer Science and Applications*, Vol.5, No.7, 30-33, 2014.