Optimal Adaptive Neuro-Fuzzy Inference System Architecture for Time Series Forecasting With Calendar Effect

by Tarno Tarno

Submission date: 25-Jul-2022 08:22PM (UTC+0700)
Submission ID: 1875006459
File name: ve_Neuro-Fuzzy_Inference_System_Architecture_for_Time_Series.pdf (1.28M)
Word count: 10942
Character count: 48485

Sains Malaysiana 51(3)(2022): 895-909
http://doi.org/10.17576/jsm-2022-5103-23

Optimal Adaptive Neuro-Fuzzy Inference System Architecture for Time Series Forecasting with Calendar Effect

(Seni Bina Sistem Inferens Neuro-Kabur Adaptif Optimum untuk Ramalan Siri Masa dengan Kesan Kalendar)

PUTRIAJI HENDIKAWATI[1,2,*], SUBANAR[1], ABDURAKHMAN[1] & TARNO[3]

[1]*Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia*
[2]*Department of Mathematics, Universitas Negeri Semarang, Semarang, Indonesia*
[3]*Department of Statistics, Universitas Diponegoro, Semarang, Indonesia*

Received: 19 January 2021/Accepted: 13 August 2021

ABSTRACT

INTRODUCTION

In recent times, the development of forecasting methods has been widely used and benefits various fields. The use of nonlinear models with the help of machine learning for forecasts has also been widely studied. Adaptive Neuro-Fuzzy Inference System (ANFIS) combines two soft computing methods, namely ANN and fuzzy logic (Jang 1993). In ANFIS, the fuzzy inference system is

implemented in the adaptive network framework. ANFIS has several advantages: a high convergence rate, good stability, a repeatable training process, high prediction precision, and very suitable for dealing with time series prediction problems (Liu & Zhou 2017). There have been many studies related to the advantages of ANFIS for prediction and forecasting, among them Duan et al. (2019), Lei and Wan (2012), Nayak et al. (2004), Sumithira and Nirmal (2014), and Wei et al. (2011). The development of ANFIS with various other methods that produce hybrid methods to get better results has also been studied by Gunasekaran and Ramaswami (2014), Liu and Zhou (2017), and Sood et al. (2020). In addition, several studies on hybrid models, including Kamisan et al. (2018) and Suhartono et al. (2019), also show that the hybrid model can give good results.

Various modelling problems in the real world are generally influenced by many potential inputs that can be incorporated into the built model. Therefore, an investigation is needed to determine the appropriate potential input that is made a priority. There is no definite procedure for choosing an ANFIS architecture that combines input variables, number of MFs, and ANFIS rules to find the optimal ANFIS. In general, there is a trial and error to find the input variable and the number of MFs. There is no standard method to determine this, therefore, various proposed new methods were given and carried out by several researchers. How to perform preprocessing to obtain optimal ANFIS is a topic discussed by several researchers, namely Azadeh et al. (2011), Polat (2012), and Yunos et al. (2008). How to find the best ANFIS model, i.e. how to find a combination in ANFIS architecture the number of input variables and the number of MFs has also been studied by several researchers such as Jang et al. (1997), Nauck (2000), Prasad et al. (2016), Tarno et al. (2017), and Septiarini and Musikasuwan (2018).

Many time series data relating to the economy are affected by many interventions such as government political policies, disaster events, or holidays in a long period of time. Interventions that can affect the data need to be considered so that data analysis results can be described properly. In real cases, some products and consumer behaviour patterns are related to the occurrence of holiday events that result in changes in the number of sales of a product according to the holiday events that occur. The religious holidays that occur are not always influenced by the Gregorian calendar, which routinely occurs on the same date and time for each period. This phenomenon is known as the calendar effect. Several studies on the effect of calendars on time series data include Cleveland and Delvin (1982), Hillmer (1982), Kling and Gao (2005), Liu (1980), Mills and Andrew (1995), Seyyed et al. (2005), Sullivan et al. (2001), and Vergin and McGinnis (1999).

One of the holiday events that occurred in Indonesia is Eid al-Fitr. Eid al-Fitr holidays are calculated based on the lunar calendar so that the time of occurrence in each year is constantly changing and has a forward pattern that shifts around 11 to 12 days. In this study, the effect of the Eid Al-Fitr holiday calendar on time series data was observed. For this purpose, actual data on the number of visitors to Tanjung Priok Port, the most populous Port in Indonesia influenced by the Eid al-Fitr holiday, is used.

The motivation in this research arises from the fact that no published works have examined time series data with calendar variations using ANFIS. With the holiday effect on time series observation data, the ARIMA model to determine the input variables proposed by Jang (1996) is no longer able and suitable to describe the data adequately in this study, therefore, the ARIMAX model is proposed to accommodate the calendar effect. By utilizing soft computing and the advantages of the ANFIS method, the hybrid ARIMAX ANFIS method will be applied to time series data with calendar variations. This paper aims to develop an ANFIS optimal architecture formation method proposed by Tarno et al. (2013) to determine the input and number of MFs in ANFIS architecture, especially for time series data influenced by calendar effects. This paper is organized as follows. Next section contains theoretical studies of identification methods in ANFIS of a time series affected by calendar effects and describes the ANFIS architecture. The following section describes the structure and learning rules of adaptive networks in time series with calendar effect. Subsequent section introduces the procedure proposed in this paper. Application examples of case studies are given in the next section. The last section concludes this paper by providing extensions and future directions for this work.

Materials and Methods

Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX)

Time series modeling can be done by using historical data and adding other variables that are considered to have a significant influence on the data to improve forecasting accuracy. ARIMAX model is a modification of the ARIMA

model with the addition of predictor variables. In this model, the factors affecting the response variable Z at time t are not only a function of Z variable in time but also by other independent variables at time t. In general, the shape of the ARIMAX(p,d,q) model is given by the equation

$$(1-B)^d \emptyset_p(B) Z_t = \mu + \theta_q(B) a_t + \alpha_1 X_{1t} + \cdots + \alpha_k X_{kt},$$

with Z_t response variable, $\emptyset_p$ is autoregressive parameter to-p, θ_q is moving average parameter to-q, X_{it}, $i = 1,2,\cdots$, k are the time series of exogenous variables (predictors), $\alpha_1,\cdots,\alpha_k$ = coefficient of exogenous variables, with $\emptyset_p(B) = (1-\emptyset_1 B-\cdots-\emptyset_p B^p)$ and $\theta_q(B) = (1-\theta_1 B-\cdots-\theta_q B^q)$ are AR and MA processes, respectively. In this model, Z_t and X_{it} are assumed to be stationary. ARIMAX modelling steps are generally the same as ARIMA modelling through three-stage: model identification, parameter estimation, and diagnostic checking Box et al. (2015). But in model estimation, the components of other independent variables are added to the model.

ANFIS ARCHITECTURE

ANFIS Architects consist of five layers built with three main components consecutive fuzzification, fuzzy inference systems, defuzzification. In the time series data with calendar effects, there are additional variables that can be input candidates, namely dummy variables, which indicate the calendar effect on the data. If there are p lag input variables, say Z_{t-1}, $Z_{t-2}\cdots$, Z_{t-p} and a number of i dummy variables that represent the calendar effect on D_1, D_2,$\cdots$, D_i data and one output Z_t with the number of membership functions is m, assuming the first-order Sugeno rules is as follows.

If Z_{t-1} is A_{11}, Z_{t-2} is $A_{21,}\cdots$, Z_{t-p} is A_{p1}, D_1 is $A_{(p+1)1}$, D_2 is $A_{(p+2)1}$ $\cdots$, D_i is $A_{(p+i)1}$, then

$$Z_t^{(1)} = \theta_{11}Z_{t-1} + \theta_{12}Z_{t-2} + \cdots + \theta_{1p}Z_{t-p} + \theta_{1(p+1)}D_1 + \theta_{1(p+2)}D_2 + \cdots + \theta_{1(p+i)}D_i.$$

If Z_{t-1} is A_{12}, Z_{t-2} is $A_{22,}\cdots$, Z_{t-p} is A_{p2}, D_1 is $A_{(p+1)2}$, D_2 is $A_{(p+2)2}$ $\cdots$, D_i is $A_{(p+i)2}$, then

$$Z_t^{(2)} = \theta_{21}Z_{t-1} + \theta_{22}Z_{t-2} + \cdots + \theta_{2p}Z_{t-p} + \theta_{2(p+1)}D_1 + \theta_{2(p+2)}D_2 + \cdots + \theta_{2(p+i)}D_i.$$

$\vdots$

If Z_{t-1} is A_{1m}, Z_{t-2} is $A_{2m,}\cdots$, Z_{t-p} is A_{pm}, D_1 is $A_{(p+1)m}$, D_2 is $A_{(p+2)m}$ $\cdots$, D_i is $A_{(p+i)m}$,

then $Z_t^{(m)} = \theta_{m1}Z_{t-1} + \theta_{m2}Z_{t-2} + \cdots + \theta_{mp}Z_{t-p} + \theta_{m(p+1)}D_1 + \theta_{m(p+2)}D_2 + \cdots + \theta_{m(p+i)}D_i.$

where Z_{t-k} is A_{kj} as premise parameter, while $Z_t^{(j)} = \theta_{j0}\sum_{k=1}^{p+i} \theta_{jk}Z_{t-k}$ as a consequent parameter, θ_{jk}, θ_{j0} as a linear parameter, A_{kj} as a nonlinear parameter with $j = 1, 2,\cdots$, m (rules), $k = 1, 2, \cdots, p, p+1, \cdots, p+i$.
If the firring strength for m (rules) is $Z_t^{(1)}, Z_t^{(2)},\cdots, Z_t^{(m)}$, are w_1, w_2, $\cdots$, w_m, then the output of Z_t can be expressed in the form

$$Z_t = \frac{w_1 Z_t^{(1)} + w_2 Z_t^{(2)} + \cdots + w_m Z_t^{(m)}}{w_1 + w_2 + \cdots + w_m}$$

Here, if the dummy variable calendar effects D_1, D_2, $\cdots$, D_i are expressed as $Z_{t-(p+1)}$, $Z_{t-(p+2)},\cdots, Z_{t-(p+i)}$, then the first order Sugeno rules bocome.

If Z_{t-1} is A_{11}, Z_{t-2} is $A_{21,}\cdots$, Z_{t-p} is A_{p1}, $Z_{t-(p+1)}$ is $A_{(p+1)1}$, $Z_{t-(p+2)}$ is $A_{(p+2)1}$ $\cdots$, $Z_{t-(p+i)}$ is $A_{(p+i)1}$, then $Z_t^{(1)} = \theta_{11}Z_{t-1} + \theta_{12}Z_{t-2} + \cdots + \theta_{1p}Z_{t-p} + \theta_{1(p+1)}Z_{t-(p+1)} + \theta_{1(p+2)}Z_{t-(p+2)} + \cdots + \theta_{1(p+i)}Z_{t-(p+i)}$.

$\vdots$

If Z_{t-1} is A_{1m}, Z_{t-2} is $A_{2m,}\cdots$, Z_{t-p} is A_{pm}, $Z_{t-(p+1)}$ is $A_{(p+1)m}$, $Z_{t-(p+2)}$ is $A_{(p+2)m}$ $\cdots$, $Z_{t-(p+i)}$ is $A_{(p+i)m}$, the $Z_t^{(m)} = \theta_{m1}Z_{t-1} + \theta_{m2}Z_{t-2} + \cdots + \theta_{mp}Z_{t-p} + \theta_{m(p+1)}Z_{t-(p+1)} + \theta_{m(p+2)}Z_{t-(p+2)} + \theta_{m(p+i)}Z_{t-(p+i)}$.

ANFIS architecture illustrated in Figure 1 consists of five layers (Jang et al. 1997) described below.

Layer 1 Each node in the first layer is adaptive with one activation function. The output of each node is the degree of membership value given by the input of the membership function.

$$\mu_{A_{11}} Z_{t-1}, \mu_{A_{12}} Z_{t-1}, \cdots, \mu_{A_{1m}} Z_{t-1}, \mu_{A_{21}} Z_{t-2}, \mu_{A_{22}} Z_{t-2}, \cdots,$$
$$\mu_{A_{2m}} Z_{t-2}, \cdots, \mu_{A_{p1}} Z_{t-p}, \mu_{A_{p2}} Z_{t-p,\cdots,} \mu_{A_{pm}} Z_{t-p}, \cdots,$$
$$\mu_{A_{(p+1)1}} Z_{t-(p+1)}, \mu_{A_{(p+1)2}} Z_{t-(p+1)}, \cdots, \mu_{A_{(p+1)m}} Z_{t-(p+1)}, \cdots,$$
$$\mu_{A_{(p+i)1}} Z_{t-(p+i)}, \mu_{A_{(p+i)2}} Z_{t-(p+i)}, \cdots, \mu_{A_{(p+i)m}} Z_{t-(p+i)}$$

The membership function used in this study is the Gaussian membership function (gaussmf) which can be stated as $\mu_{A_{kj}}(Z_{t-k}) = exp\left(-\frac{1}{2}\left(\frac{Z_{t-k} - c_{jk}}{a_{jk}}\right)^2\right)$, with $j = 1, 2, \cdots, m$; $k = 1, 2, \cdots, p, p+1, \cdots, p+i$. This parameter is called the premise parameter.

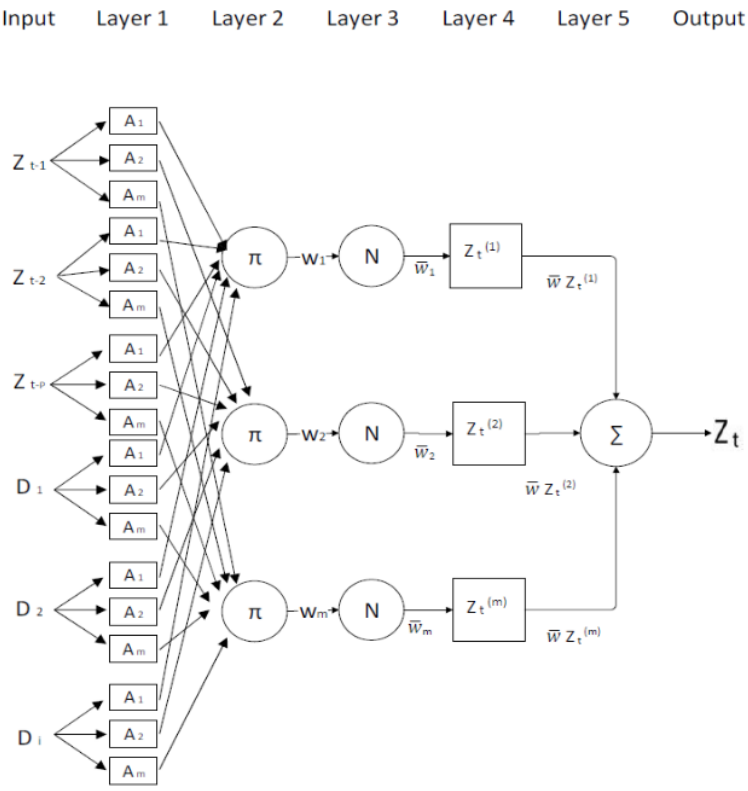


FIGURE 1. ANFIS architecture with a dummy calendar effect

Layer 2 Each node in the second layer is a fixed node where the output of this layer is the sum of the incoming signals. Generally used AND fuzzy operators. Each node represents the firing strength of w_j from rule to j-th.

$$w_j = \prod_{k=1}^{p+i} \mu_{A_{kj}}(Z_{t-k}), \qquad j = 1,2,\cdots,m$$

Layer 3 All nodes in this layer are fixed nodes, which is the result of calculating the ratio of firing strength to j-th with the sum of all the existing firing strengths of the rules.

$$\overline{w_j} = \frac{w_j}{\sum_{j=1}^{m} w_j}$$

Layer 4 Every node is an adaptive node with output for each node defined as

$$\overline{w_i}Z_t^{(j)} = \overline{w_i}(\theta_{j1}Z_{t-1} + \theta_{j2}Z_{t-2} + \cdots + \theta_{jp}Z_{t-p} + \theta_{j(p+1)}Z_{t-(p+1)} + \cdots + \theta_{j(p+i)}Z_{t-(p+i)})$$

with $j = 1, 2, \cdots, m$ and w_j is the normalized firing strength in the third layer, with $\theta_{j1}, \theta_{j2}, \cdots, \theta_{jp}, \theta_{j(p+1)}, \cdots, \theta_{j(p+i)}$ being the consequent parameters.

Layer 5 This layer produces a single node that is a fixed node that computes all incoming signals, the output is the overall output of the network.

$$Z_t = \sum_{j=1}^{m} \overline{w_i}(\theta_{j1}Z_{t-1} + \theta_{j2}Z_{t-2} + \cdots + \theta_{jp}Z_{t-p} + \theta_{j(p+1)}Z_{t-(p+1)} + \cdots + \theta_{j(p+i)}Z_{t-(p+i)})$$

We used a hybrid learning algorithm, which in forward pass the consequent parameter is identified by the least-squares method. Meanwhile, in the backward pass, the premise parameter is updated using gradient descent.

Based on architecture with these five layers, the general model of ANFIS can be expressed as

$$Z_t = \sum_{j=1}^{m}\sum_{k=1}^{p+i} \theta_{jk}\left(\overline{w_j}Z_{t-k}\right) + \sum_{j=1}^{m} \theta_{j0}\,\overline{w_j}$$

LAGRANGE MULTIPLIER TEST PROCEDURE FOR ADDING VARIABLE

Lagrange Multiplier (LM) test is used to test hypotheses related to adding variables and the number of membership functions to the ANFIS architecture. The determination of input variables using LM test procedure, begins with testing using a minimum number of inputs, number of membership functions, and rules. The first stage, the ANFIS model was formed by using 1 input variable selected from several input candidates, 2 number of membership functions, and 2 rules. The variable, which was first tested to be included in the ANFIS architecture, was the variable with the largest R^2 value from the previous partial test.

In data with calendar effects, additional variables can be input candidates, namely dummy variables that refer to the calendar effect on the data. For p lag input variables, say Z_{t-1}, Z_{t-2} ..., Z_{t-p} and a number of i dummy variables that state the calendar effect symbolized by $D_1, D_2, \cdots, Di$ with the number of MFs of m, then the restricted model for this case can be stated as

$$Z_t = \sum_{j=1}^{m}\sum_{k=1}^{p+i} \theta_{jk}\left(\overline{w_j}Z_{t-k}\right) + \sum_{j=1}^{m} \theta_{j0}\,\overline{w_j} + \varepsilon_t$$

where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ and unrestricted model to add one input $Z_{t-(p+i+1)}$ is

$$Z_t = \sum_{j=1}^{m}\sum_{k=1}^{p+i+1} \theta_{jk}\left(\overline{w_j}Z_{t-k}\right) + \sum_{j=1}^{m} \theta_{j0}\,\overline{w_j} + \upsilon_t$$

where $\upsilon_t \sim N(0, \sigma_\upsilon^2)$.

The null hypothesis for testing the addition of variables is formulated as follows,

$$H_0: \theta_{1(p+i+1)} = \theta_{2(p+i+1)} = \cdots = \theta_{m(p+i+1)} = 0$$

If the $LM = nR_{\varepsilon_t}^2 > \chi^2_{(\alpha,df)}$ then H_0 is rejected.

The LM test introduced by Lee et al. (1993) and Terasvirta et al. (1994) was also used to test linearity. The test is carried out through the χ^2 test with the following procedure (Gujarati 2009).

i. Regress Z_t to Z_{t-1}, Z_{t-2}, ⋯, Z_{t-p}, $Z_{t-(p+1)}$,⋯, $Z_{t-(p+i)}$ and estimate the parameters on the restricted model using the OLS method.

ii. Calculate the residual estimated $\hat{\varepsilon}_t$ from the regression, with

$$\hat{\varepsilon}_t = Z_t - \sum_{j=1}^{m}\sum_{k=1}^{p+i} \theta_{jk}\left(\overline{w_j}Z_{t-k}\right) - \sum_{j=1}^{m} \theta_{j0}\,\overline{w_j}$$

Regress $\hat{\varepsilon}_t$ to Z_{t-1}, Z_{t-2}, ⋯, Z_{t-p}, $Z_{t-(p+1)}$,⋯, $Z_{t-(p+i)}$ and m additional predictors, then calculate the coefficient of determination R^2 from the regression.

THE PROCEDURE OF THE PROPOSED METHOD: ARIMAX-ANFIS BASED THE LM TEST

The determination of ANFIS input in time series cases can be identified by the significant lag partial autocorrelation function (PACF) plot. ARIMA subset model can be formed based on significant lag, which is then used to model time series data affected by the Eid al-Fitr holidays. The subset ARIMA model can be easily identified, estimated, and used for forecasting by forming a representative parsimony model. Furthermore, the subset ARIMA develops into the ARIMAX by adding a calendar effect variable as an exogenous variable.

At this stage, the variable that refers to the Eid al-Fitr holiday calendar effect intervention is defined. Forming a calendar effect variable is done in the following two ways.

a. Dummy variable calendar effect

The first way is to use a dummy variable to declare the Eid al-Fitr holiday.

$$D_{t-1} = \begin{cases} 1 & month\ before\ Eid\ al\_Fitr \\ 0 & other\ month \end{cases}$$

$$D_t = \begin{cases} 1 & month\ of\ Eid\ al\_Fitr \\ 0 & other\ month \end{cases}$$

$$D_{t+1} = \begin{cases} 1 & month\ after\ Eid\ al\_Fitr \\ 0 & other\ month \end{cases}$$

In this model, the intercept is removed to avoid the dummy variable trap.

b. Variable days proportion calendar effects

The second method is done by calculating day proportions by assuming Eid al-Fitr events are distributed uniformly over 10 days, starting from 3 days before Eid and 7 days after that, including Eid al-Fitr (Liu 1986). The days proportion calendar effect DP_t are set as follows and shown in Table 1.

TABLE 1. The days proportion of the Eid al-Fitr calendar effect

Year	Date	Month	Week	DP_{t-1}	DP_t	DP_{t+1}
2006	24	10	4	0	1	0
2007	13	10	2	0	1	0
2008	1	10	1	0.3	0.7	0
2009	20	9	3	0	1	0
2010	9	9	2	0	1	0
2011	30	8	4	0	0.5	0.5
2012	18	8	3	0	1	0
2013	7	8	3	0	1	0
2014	28	7	4	0	0.7	0.3
2015	17	7	3	0	1	0
2016	6	7	1	0	1	0
2017	25	6	4	0	0.9	0.1
2018	15	6	3	0	1	0
2019	5	6	1	0	1	0

THE PROPOSED PROCEDURE

Modeling problems in the real world are generally influenced by many potential inputs that can be incorporated into the built model. Therefore, an investigation is needed to determine the appropriate potential input that is made a priority. This study constructs the ANFIS architecture with preprocessing stages using ARIMAX and the LM test inference procedure. The LM test is used to test hypotheses for the determination of input variables and the number of membership functions to form the optimal ANFIS architecture for prediction of time series, which is affected by calendar effects. The data analysis steps in this study are as follows.

PREPROCESSING DATA

Input determination begins by plotting a PACF plot from time series data. The PACF plot is used to identify whether a lag variable affects the data. If the PACF lag value to $k(\emptyset_{kk})$ twice the standard error $\emptyset_{kk}$, then the

lag-*k* can be identified as an ANFIS input variable. Based on the significant lag and calendar effect, the ARIMAX model can be identified. This model contains several input variables that can be entered into the ANFIS model. Furthermore, the formation of the ARIMAX model is continued with parameter estimation to see significant variables, and diagnostic testing is carried out by testing the residual independence and normality. The ARIMAX model that meets all the conditions with the smallest Akaike Information Criterion (AIC) value is then determined as a model of the ARIMAX calendar effect.

FORECASTING WITH ANFIS

Determine the appropriate input variables

The first input to be entered in the model is determined based on the R^2 value. The variable with the largest R^2 will be the first input. Determination of variable inputs in the ANFIS model is done one by one on all existing input candidates until all suitable input candidates are tested. At this stage, all input variables that will meet the LM test will be obtained. The optimization of variable input stops when the LM test value is not significant for the addition of input. At this stage, taking into account the principle of parsimony, ANFIS architecture is used with two membership functions and two rules.

The steps for using the LM Test to determine the input variable in accordance with the subsection *LM Test Procedure for Adding Variable* are described as follows.

i. Choose the first input variable that has the largest R^2.

ii. Estimating parameters in the restricted ANFIS model with the output variable Z_t.

Suppose the first input variable is Z_{t-1}, then $Z_t = \overline{w_1}(\hat{\theta}_{11}Z_{t-1} + \hat{\theta}_{10}) + \overline{w_2}(\hat{\theta}_{21}Z_{t-1} + \hat{\theta}_{20}) + \varepsilon_t$.

iii. Calculates the estimated residual (ε_t) value of the restricted model.

iv. Enter an additional candidate variable input that is the candidate input variable with the next largest R^2 value, supposed Z_{t-2}.

v. Form an unrestricted ANFIS model to increase the number of input lag variables (input becomes 2) with the residuals of the restricted model being input variables

$$\varepsilon_t = \overline{w_1}(\hat{\theta}_{11}Z_{t-1} + \hat{\theta}_{12}Z_{t-1} + \hat{\theta}_{10}) + \overline{w_2}(\hat{\theta}_{21}Z_{t-2} + \hat{\theta}_{22}Z_{t-2} + \hat{\theta}_{20}) + \upsilon_t.$$

vi. Calculates the value of $R^2_{\varepsilon_t}$ from the regression estimation of residual ε_t values and unrestricted ANFIS models for the addition of one input lag.

vii. Determine the conclusions of the hypothesis LM test.

Repeat steps (i) to (vii) until all the candidate input variables obtained from the best ARIMAX model are all tested. Furthermore, increasing the number of input variables continues so that all the inputs variables can be determined.

Determine the optimal number of membership functions.

The steps for using the Lagrange Multiplier Test to determine the number of membership functions is as follows.

i. Forming the ANFIS model by entering all the input variables selected in the previous stage by increasing the number of clusters starting from 2 clusters then calculating the RMSE and MAPE value of the ANFIS architecture that was formed.

ii. Increase the number of membership functions to the optimal number of membership functions that provide the smallest RMSE and MAPE value.

iii. Determine the optimal number of membership functions that provide the smallest RMSE and MAPE using the LM test previously described.

Forecasting

Forecasting is done using the ANFIS architecture from the results in *Determine the optimal number of membership functions* and *Forecasting* steps.

i. The initial stage is done by determining the input and output variables.

ii. Divide the data into two parts, namely training (insample) and testing (outsample).

iii. Determine the Gauss function as membership function.

iv. Use the number of MFs that satisfy the LM test obtained in *Forecasting* steps.

v. Training ANFIS parameters with the training

and testing data. At this stage, the RMSE and MAPE values for the training and testing process for all ANFIS architectures will be obtained. The best ANFIS architecture is determined by looking at the smallest RMSE and MAPE in the testing data.

The procedure of the method proposed in this study is

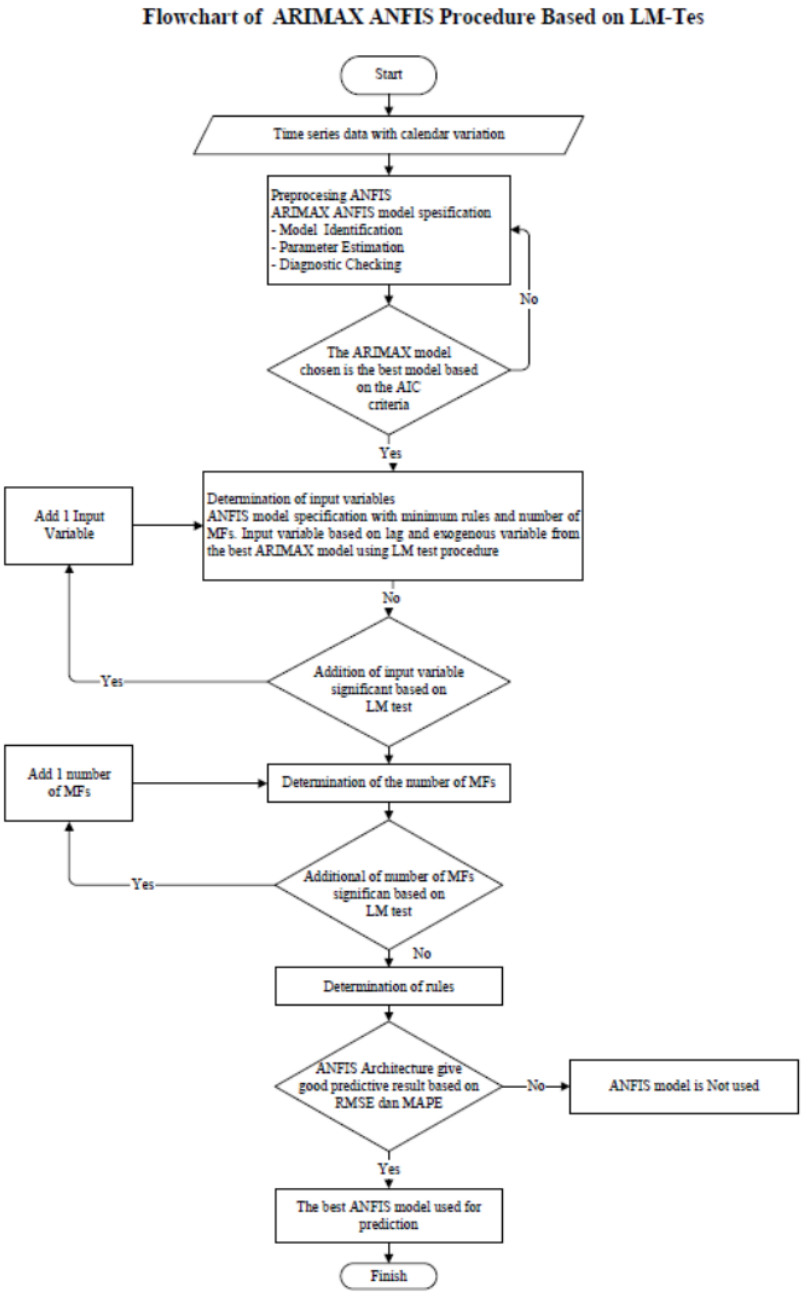


FIGURE 2. The procedure for the proposed method

illustrated in Figure 2.

ACCURACY CRITERIA

There are many criteria that can be used to evaluate forecasting methods, the accuracy of forecasting is generally the basis for determining the appropriate model. Measurement error forecasting accuracy has been widely studied by experts to investigate the accuracy of various forecasting methods (Makridakis et al. 1997). There are three performance measurements used in this study to evaluate the accuracy of the proposed methods both in training and testing data, namely Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and Coefficient of determination (R^2). These three measurement criteria are stated with,

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \left(Z_t - \hat{Z}_t\right)^2}$$

$$MAPE = 100 \times \frac{1}{n}\sum_{t=1}^{n} \left|\frac{Z_t - \hat{Z}_t}{Z_t}\right|$$

$$R^2 = 1 - \frac{\sum_t (Z_t - \hat{Z}_t)^2}{\sum_t (Z_t - \bar{Z}_t)^2}$$

RESULTS AND DISCUSSION

This paper's data study is the monthly volume of visitors to Jakarta's Tanjung Priok Port from January 2006 to November 2019 obtained from Statistics Indonesia. By examining the data, the Eid holiday causes an increase in the recurring pattern in the months of Eid al-Fitr each year. Figure 3(a) illustrates this pattern. Based on the

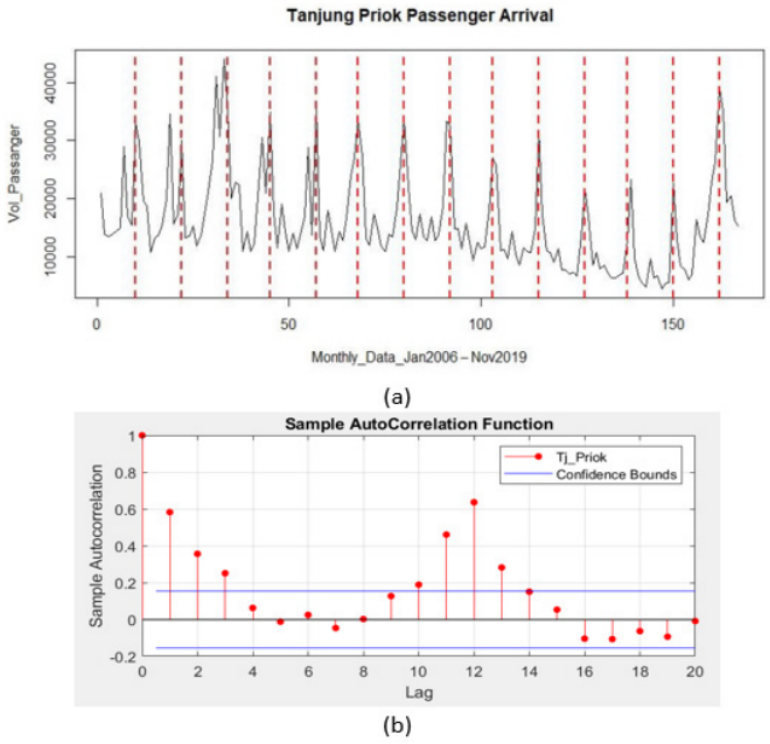


FIGURE 3. (a) Tanjung Priok Port Passenger Plot, (b) Sample PACF Plot

data plot, the data has a recurring pattern every 11 or 12 months. There is a pattern that is almost the same in every data point with a red line. During this time, there was a pattern of a significant increase in data. This pattern indicates a calendar effect on the data.

After determining the calendar effect dummy variable (Table 1), then at the data preprocessing stage the input variable was determined by applying the ARIMAX and LM test models. This process begins with modeling the data using ARIMA to see the significant lag and the most efficient model. In this study, for convenience and simplicity, the ARIMA model used is limited to only using lag data from the AR section and does not take into account the lag in the MA section. This is done by only paying attention to the PACF plot of the observed data. Based on the PACF plot in Figure 3(b), it can be seen that the significant lag is lag 1, 2, 3, 11,12, 13. From Table 2, based on the AIC model value, significant lag, and the number of variables that affect the model, based on the parsimony principle, the ARIMA([1,11,12],1,0) model chosen as the best model used for the next stage.

TABLE 2. Determining significant input variable using ARIMA

Model	ARIMA ([1,11,12],1,0)	ARIMA ([11,12],1,0)	ARIMA ([11,12,13],1,0)	ARIMA ([1,2,3,11,12,13],1,0)	ARIMA ([1,11,12,13],1,0)	ARIMA ([1,2,11,12,13],1,0)
a_1	*Sig*			*Sig*	*Sig*	*Sig*
a_2						*Sig*
a_3				*Not Sig*		
a_{11}	*Sig*	*Sig*	*Sig*	*Sig*	*Sig*	*Sig*
a_{12}	*Sig*	*Sig*	*Sig*	*Sig*	*Sig*	*Sig*
a_{13}			*Sig*	*Sig*	*Sig*	*Sig*
R^2	0.63	0.56	0.57	0.67	0.67	0.67
SSR	3.35E+09	4.16E+09	4.05 E+09	3.05E+09	3.14E+09	3.08 E+09
AIC	19.763	19.984	19.964	19.729	19.738	19.725
White Noise						
Num of Variable	3	2	3	6	4	5

Furthermore, calendar effect variables that significantly affect the data will also be candidates for input variables in ANFIS. Table 3 shows the calendar effect ARIMAX model which contains significant variables as an alternative to determining the input variables for the ANFIS model. There are six candidate input variables that can be included in the ANFIS model. The R^2 for each variable are 0.396 for *Lag-1,* 0.255 for *Lag-11,* 0.478 for *Lag-12*, 0.259 for D_t 0.020 for D_{t+1} and 0.259 for DP_t. The following Table 4 is the result of testing the addition of variable input in the Tanjung Priok Port Passenger data. Based on the value of R^2, *Lag-12* is selected as the first input variable of the ANFIS architecture. The next step is to add other inputs to the ANFIS architecture in stages according to the large order value of R^2. The step-by-step results from the analysis of adding input variables are shown in Table 4. The LM test value of the three models shows a value greater than the $\chi^2_{(\alpha,\, df)}$ which means that additional input variables can be accepted and included in the model. We can conclude that the optimal input variables for the calendar effect of dummy models are *Lag-1*, *Lag-11*, *Lag-12*, D_t, and D_{t+1}. Meanwhile, the optimal input variables for the days proportion calendar effect models are *Lag-1*, *Lag-11*, *Lag-12*, and DP_t. This shows that, all significant input candidates based on the ARIMAX model can be entered into ANFIS input variables.

TABLE 3. Forecasting calendar effect using ARIMAX

Model	Input Variable	R^2	SSR	AIC	White Noise	Normality Residual
ARIMA([1,11,12],1,0)	*Lag-1*	0.63	3.45E+09	19.806	√	×
	Lag-11					
	Lag-12					
ARIMA([1,11,12],1,0)	*Lag-1*	0.64	3.35E+09	19.789	√	√
with D_t, D_{t+1},	*Lag-11*					
	Lag-12					
	D_t					
	D_{t+1}					
ARIMA([1,11,12],1,0)	*Lag-1*	0.66	3.25E+09	19.763	√	√
with DP_t	*Lag-11*					
	Lag-12					
	DP_t					

TABLE 4. ANFIS variable input determination

Model	Input Variable	RMSE	LM Stat	Conclusion
Without dummy variables calendar effect				
ARIMA([1,11,12],1,0)	*Lag-12*	5982.4	68.288	var added
	Lag-12, Lag-1	5509.5	31.072	var added
	Lag-12, Lag-1, Lag-11	5374.9	36.421	var added
Dummy variables calendar effect				
ARIMA([1,11,12],1,0)	*Lag-12*	5982.4	68.295	var added
with D_t, D_{t+1},	*Lag-12, Lag-1*	5509.5	31.106	var added
	Lag-12, Lag-1, D_t	5115.0	46.542	var added
	Lag-12, Lag-1, D_t*,Lag-11*	5012.0	50.444	var added
	Lag-12, Lag-1, D_t *, Lag-11,* D_{t+1}	4559.6	66.411	var added
Days proportion calendar effect variable				
ARIMA([1,11,12],1,0)	*Lag-12*	5982.4	68.302	var added
with DP_t	*Lag-12, Lag-1*	5509.6	31.076	var added
	Lag-12, Lag-1, DP_t	5084.0	46.712	var added
	Lag-12, Lag-1, DP_t*, Lag-11*	5022.9	50.000	var added

After obtaining the input variable, the next step is to determine the number of MFs. The number of MFs used starts from 2 and then gradually increases until the maximum number of MFs gives the smallest error value. In determining the optimum number of MFs, a fuzzy C-Means (FCM) clustering technique is used, which is determined by using the LM test with the procedure described earlier.

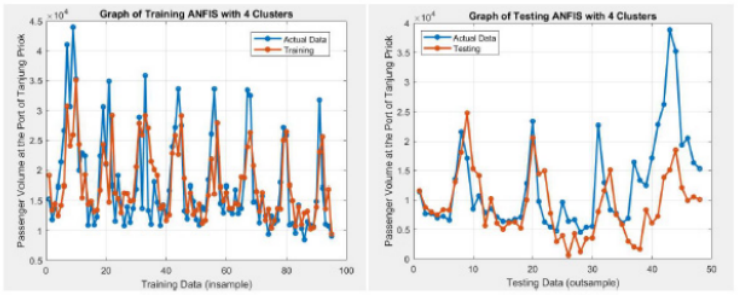
The number of MFs is restricted so that the estimated number of parameters is not more than the amount of data analyzed so that the resulting error tends to increase. Because the number of parameters estimated does not more than the observational data, 3 are the maximum number of MFs that can be used for models with calendar effects and 4 number of MFs for models without including calendar effects. When using to many MFs, the results obtained may be better, but there is a danger that if too many membership functions are used, the system will become overfitted. Overfitting can make the prediction results too precise for the training data, and therefore that does not give good results on other data (testing).

Table 5 shows that the ANFIS models can use 2 to 4 numbers of MFs because of the LM test value more significant than the $\chi^2_{(a,\ df)}$. For models with calendar effect variables, using 3 number of MFs give the smallest RMSE. At this stage, a significant input variable has been obtained and the optimal number of MFs for the ANFIS architecture will be used for forecasting. Forecasting is carried out using the ANFIS architecture obtained in the previous stage. First, divide the data into two parts: data training (in sample) from January 2006 to December 2016 and data testing (out sample) from January 2017 to November 2019. The input nodes are the previous lagged observation that is significant to the data based on the results of preprocessing data with ARIMAX. At the same time, the output provides the forecasting for future values. The previous lagged that uses as an input variable is a significant lag. The ANFIS architecture model used is a model with a significant input variable and the optimal number of MF based on the LM test. As a limitation, the membership function used in ANFIS is Gaussian. Gaussian chose because of its simple function, with only two parameters (mean and variance) estimated.

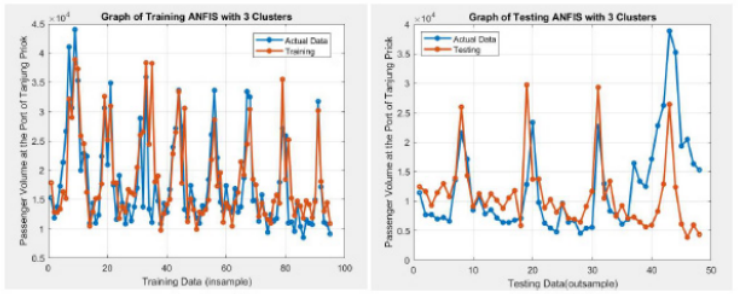
TABLE 5. Determination of the number of ANFIS membership function

Num of MFs	RMSE	LM Stat	Conclusion
Without dummy variables calendar effect with 3 input variables			
2	6228.4	68.316	MFs can be added
3	5097.9	11.227	MFs can be added
4	4977.8	17.817	MFs can be added
Dummy variables calendar effect with 5 input variables			
2	4549.2	103.129	MFs can be added
3	4412.8	9.370	MFs can be added
Days proportion calendar effect variable with 4 input variables			
2	5013.4	94.657	MFs can be added
3	4705.6	20.840	MFs can be added

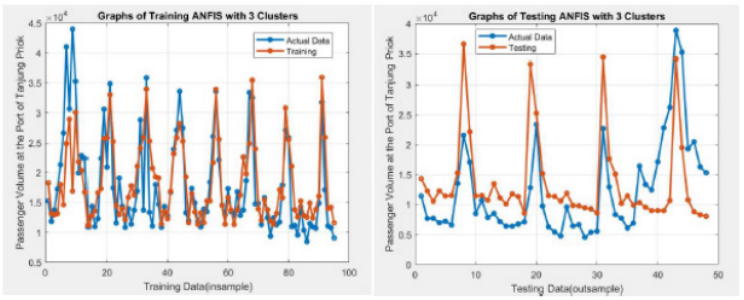
Table 6 summarizes the results of ANFIS forecasting in the training and testing stages by using an optimal variable input and numbers of MFs from the previous step. In a fuzzy system, each number of MFs is considered a rule. Therefore, the number of fuzzy rules is equal to the number of membership functions developed with FCM.



(a) Without Dummy Variable Calendar Effect



(b) With Dummy Variable Calendar Effect



(c) With Dummy Proportional Variable Calendar Effect

FIGURE 4. Plot training and testing ANFIS

TABLE 6. Training and testing ANFIS

Number of MFs	RMSE		MAPE		R^2	
	Training	Testing	Training	Testing	Training	Testing
Without dummy variables calendar effect with 3 input variables						
4	6147.9	9119.9	9.362	14.156	0.453	0.698
Dummy variables calendar effect with 5 input variables						
3	4948.3	7799.0	7.297	13.021	0.661	0.882
Days proportion calendar effect variable with 4 input variables						
3	5633.6	8437.7	8.469	13.829	0.635	0.768

In this case, the ANFIS architecture is tested with the best number of MF used based on the previous stage. In the calendar effect dummy model used, the use of several numbers of MFs affects the forecast error value. This shows that increasing the number of MFs to a certain amount can improve accuracy. However, if the number of MFs is used too much, it will make the analysis process longer because more parameters must be estimated. Based on results in Table 6, the optimal architecture for both the dummy and the days proportion calendar effect model is obtained when using three numbers of MFs with five and four input variables, respectively. The RMSE testing value for both models was 7799.0 and 8437.7, respectively, with MAPE testing being 13.021 and 13.829. The testing error use as an accurate measure of the performance model. Therefore, the best model occurs when the testing error is minimal.

In constructing time series models, it is generally assumed that interventions have effect on the overall data pattern. In this study, data on the volume of passengers at Tanjung Priok Port is influenced by calendar variations, namely the effect of the Eid al-Fitr holiday. The shape of the calendar effect patterns that occur on the data during the Eid al-Fitr holiday is seen to have a linear increase. The ARIMAX model represents the calendar effect in the data. By combining the ARIMA model and calendar effect, parameter estimation is obtained. As a comparison, ARIMA model is also formed without a calendar effect. It seems that a model that considers calendar effects provides less error than a model without including calendar effects in the analysis. We can see this from the RMSE and MAPE values of the ARIMA model, which are higher than the ARIMAX model and the smaller coefficient of determination compared to models that include calendar effects. Figure 4 shows an illustration of forecast training and test data with ANFIS optimal architecture. In this study, it has been shown that calendar interventions can significantly influence the data patterns. When there is a calendar effect, an initial approach to the data is needed before identifying the model. This paper presents a comprehensive step for identifying and estimating time series models affected by calendar interventions.

Conclusion

The proposed method for selecting input and determining the number of MF for ANFIS uses the ARIMAX model, and the LM test is tested on real-world problems; the number of visitors to Tanjung Priok Port that influenced by the calendar effects. ARIMAX can capture the effect of calendar variations on time series data. Based on the result, LM tests can be considered as an alternative way to determine input variables and number of MFs in ANFIS. Variables that are known to have no significant effect from the beginning have been eliminated so that the possibility of using too many input variables but no significant impact can be minimized. In the time series, data indicated to be influenced by calendar effects, ANFIS training and testing results indicate that for predicting time series data by entering the calendar effect gives better results when compared to without entering the calendar effect variable in the calculation. The small RMSE and MAPE values indicate this. The use of two types of calendar effect variables in this study shows that using the dummy calendar effect provides more accurate results than the days proportion calendar effect. The proposed method can be an alternative way that provides effective results for determining input priorities for ANFIS modelling.

Acknowledgements

This research was supported by Indonesia Endowment Fund for Education (LPDP) under the Doctoral Fellowship BUDI-DN. The authors thank the reviewers and editor for their valuable comments and suggestions that significantly improved the initial manuscript.

References

Optimal Adaptive Neuro-Fuzzy Inference System Architecture for Time Series Forecasting With Calendar Effect

ORIGINALITY REPORT

12%	6%	10%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	P Hendikawati, Subanar, Abdurakhman, Tarno. "Hybrid ARIMAX-ANFIS based on LM Test for Prediction of Time Series with Holiday Effect", Journal of Physics: Conference Series, 2021 Publication	4%
2	www.ssa.gov Internet Source	1%
3	Tarno, A Rusgiyono, Sugito. "Adaptive Neuro Fuzzy Inference System (ANFIS) approach for modeling paddy production data in Central Java", Journal of Physics: Conference Series, 2019 Publication	<1%
4	repository.usu.ac.id Internet Source	<1%
5	rdo.psu.ac.th Internet Source	<1%
6	www.coomeva.com.co Internet Source	<1%

7	www.iccat.int Internet Source	<1 %
8	Submitted to Cranfield University Student Paper	<1 %
9	apps.itd.idaho.gov Internet Source	<1 %
10	lawrence.edu Internet Source	<1 %
11	www.cra.gov.co Internet Source	<1 %
12	Submitted to Curtin University of Technology Student Paper	<1 %
13	ldfebui.org Internet Source	<1 %
14	www.swebowl.se Internet Source	<1 %
15	repository.tudelft.nl Internet Source	<1 %
16	Submitted to Flinders University Student Paper	<1 %
17	Nana Zhang, Kun Zhu, Shi Ying, Xu Wang. "KAEA: A Novel Three-stage Ensemble Model for Software Defect Prediction", Computers, Materials & Continua, 2020 Publication	<1 %

18	www.science.gov Internet Source	<1%
19	Ulubasoglu, M.A.. "International comparisons of rural-urban educational attainment: Data and determinants", European Economic Review, 200710 Publication	<1%
20	www.dirittobancario.it Internet Source	<1%
21	semioffice.com Internet Source	<1%
22	www.teaneckschools.org Internet Source	<1%
23	Suparta, Wayan, and Kemal Maulana Alhasa. "Adaptive Neuro-Fuzzy Interference System", SpringerBriefs in Meteorology, 2016. Publication	<1%
24	Submitted to University of Bristol Student Paper	<1%
25	Submitted to University of Sydney Student Paper	<1%
26	Xiaoyong Liu, Zhili Zhou. "A novel prediction model based on particle swarm optimization and adaptive neuro-fuzzy inference system", Journal of Intelligent & Fuzzy Systems, 2017 Publication	<1%

27	archive.org Internet Source	<1%
28	www.mass.gov Internet Source	<1%
29	Hiremath, Gourishankar S., and Jyoti Kumari. "Is There Long Memory in Indian Stock Market Returns? An Empirical Search", Journal of Asia-Pacific Business, 2015. Publication	<1%
30	Submitted to Monash University Student Paper	<1%
31	ijsrset.com Internet Source	<1%
32	www.unifr.ch Internet Source	<1%
33	"Soft Computing in Data Science", Springer Science and Business Media LLC, 2019 Publication	<1%
34	www.researchgate.net Internet Source	<1%
35	B KORKIE. "A clinical analysis of a professionally managed portfolio", Performance Measurement in Finance, 2002 Publication	<1%

36	Choo Huck Ooi, Mohamad Adam Bujang, Tg Mohd Ikhwan Tg Abu Bakar Sidik, Romano Ngui, Yvonne Ai-Lian Lim. "Over two decades of Plasmodium knowlesi infections in Sarawak: Trend and forecast", Acta Tropica, 2017 Publication	<1 %
37	porto.polito.it Internet Source	<1 %
38	Donghoon Baek, Ju-Hwan Seo, Joonhwan Kim, Dong-Soo Kwon. "Hysteresis Compensator with Learning-based Pose Estimation for a Flexible Endoscopic Surgery Robot", 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019 Publication	<1 %
39	Tarno Tarno, Suparti Suparti, Dwi Ispriyanti. "Modeling cayenne production data in Central Java using adaptive neuro fuzzy inference system (ANFIS)", Model Assisted Statistics and Applications, 2018 Publication	<1 %
40	Submitted to Università Carlo Cattaneo - LIUC Student Paper	<1 %
41	Xueyang Zhang, Jianqiang Wang, Hongyu Zhang, Junhua Hu. "A Heterogeneous Linguistic MAGDM Framework to Classroom	<1 %

	Teaching Quality Evaluation", EURASIA Journal of Mathematics, Science and Technology Education, 2017 Publication	
42	auzefkitap.istanbul.edu.tr Internet Source	<1 %
43	cpb-us-w2.wpmucdn.com Internet Source	<1 %
44	dspace.lboro.ac.uk Internet Source	<1 %
45	tind-customer-agecon.s3.amazonaws.com Internet Source	<1 %
46	www.zora.uzh.ch Internet Source	<1 %
47	Zhong Tan. "Propagation of density-oscillations in solutions to the compressible Navier-Stokes-Poisson system", Chinese Annals of Mathematics Series B, 09/2008 Publication	<1 %
48	Ahmed S. Mubarak, Hamada Esmaiel, Ehab Mahmoud Mohamed. "LTE/Wi-Fi/mmWave RAN-Level InterworkingUsing 2C/U Plane Splitting for Future 5GNetworks", IEEE Access, 2018 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

Optimal Adaptive Neuro-Fuzzy Inference System Architecture for Time Series Forecasting With Calendar Effect

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15