Hybrid ARIMAX-ANFIS Based on LM Test For Prediction Of Time Series With Holiday Effect

by Tarno Tarno

PAPER · OPEN ACCESS

Hybrid ARIMAX-ANFIS based on LM Test for Prediction of Time Series with Holiday Effect

To cite this article: P Hendikawati *et al* 2021 *J. Phys.: Conf. Ser.* **1863** 012061

View the article online for updates and enhancements.

Hybrid ARIMAX-ANFIS based on LM Test for Prediction of Time Series with Holiday Effect

P Hendikawati[1*], Subanar[2], Abdurakhman[2], Tarno[3]

[1]Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Negeri Semarang, Indonesia
[2]Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Gadjah Mada, Indonesia
[3]Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Diponegoro, Indonesia

*E-mail: putriaji.mat@mail.unnes.ac.id

Abstract.

1. Introduction

This paper highlights problems that can occur in the monthly time series. In Indonesia, economic activity changes in a unique pattern, whether to increase or decrease over certain times. The economic activity pattern ahead of religious holidays, such as Eid al-Fitr, has generally increased significantly. We can categorize the Eid holiday pattern as the occurrence of a calendar effect. Calendar effect on time series model different with seasonal adjustment. ARIMA Box Jenkins, a model that has been widely used and recognized for its success in time series forecasting, is not sufficient to capture calendar variation because they are not a precise period. Furthermore, to deal with calendar variations, ARIMA developed into ARIMAX by adding calendar effect variables as exogenous variables. Several studies of time series forecasting by applying ARIMAX include [1], [2], and [3]. In this study, the ARIMAX model was used in the preprocessing stage to describes the calendar effect in time series data.

Currently, many hybrid methods for prediction have been developed. Adaptive Neuro Fuzzy Inference System (ANFIS) is a soft computing method that has been widely used for time series data analysis. ANFIS is a combination of ANN and fuzzy logic that combines the advantages of both methods [4]. ANFIS and fuzzy logic has also been widely applied in time series data, as has been researched [5],

[6], [7], [8], [9], and [10]. Several ANFIS hybrid methods for forecasting have also been applied by [11], [12], [13], [14] and show promising results. This study proposes a hybrid ARIMAX-ANFIS model to improve prediction accuracy.

The determination of ANFIS input in time series can be identified by the lag of the partial autocorrelation function (PACF) plot. The plot lag can identify input autoregressive (AR) and see the lag and the observational data's linearity. Based on a significant lag and PACF plot, we can obtain the candidate input variables for ANFIS architecture. ARIMAX-ANFIS offered to capture calendar effects that occur on economic data in Indonesia. In M4 Competition, Makridakis [15] found that the hybrid method can be considered as an approach to increase forecasting accuracy. The hybrid between ARIMAX as a linear model and ANFIS as a non-linear model aims to capture and exist calendar effects and produce a model that can predict accurately. The use of ARIMAX-ANFIS for data with calendar effects has previously been carried out by [16], which shows promising results.

This paper applied the hybrid ARIMAX-ANFIS by applying a Lagrange Multiplier (LM) test to determine the optimal ANFIS architecture. LM test was used to test hypotheses related to adding variables and the number of membership functions to the ANFIS architecture. This research is a development from [17] previous research. The purpose of this research is to provide empirical evidence on the ARIMAX-ANFIS for time series contain a calendar effect. We conduct an empirical study with simulation and real data. This study will focus on two cases: with and without recognizing possible calendar variations in the data, in this case, the holiday effect.

The Ramadhan and Eid al-Fitr are usually special moments for most people in a Muslim country like Indonesia. From an economic point of view, in the years before the 2019 Corona Virus Disease pandemic, the Ramadan and Eid periods became the peak momentum for national economic growth. Public spending activities increase drastically every time during the Ramadan fasting period until Eid arrives. [18] reports that the purchase of foodstuffs during this period has increased. Apart from the main foodstuffs, other goods that are also experiencing demand include the textile/clothing industry, vehicle rental services, public transportation tickets, land, sea, and air [19]. The need for food and drinks before Eid continues to increase amid the Covid-19 pandemic [20]. Currently, there is a significant increase in demand, which ultimately increases the price of goods. This is what underlies the researcher to analyze the foodstuff price index data. The effect of the calendar also felt in another country, especially in the economic field. The calendar effect in economics has received attention from several researchers [21], [22], [23], [24], [25], [26], and [27].

We expected this research would provide information and recommendations to analyze data with holiday effects. The rest of this paper is organized as follows: in section 2, we describe the methods proposed in this paper. Section 3 presents results and discussion from simulation and empirical study with calendar effects and compares performance accuracy. In section 4, we draw our conclusions.

2. Methods

This study builds ANFIS architecture, starting from determining the input variable and the number of membership functions by applying ARIMAX and LM test inference procedures. In data with calendar intervention, the holiday effect can be additional variables that are stated as dummy variables. The data analysis steps in this study are as follows, as illustrated in Figure 1.

a. Preprocessing data

Input determination begins by plotting a partial autocorrelation function (PACF) from time series data. The PACF plot is used to identify whether a lag variable significantly affects the data. At this stage, the intervention of the Eid al-Fitr holiday calendar effect defined in a dummy variable that declares before (D_{t-1}), during (D_t), and after Eid al-Fitr (D_{t+1}) holidays. ARIMAX then models several potential input variable candidates. The ARIMAX model that meets the test requirements, which has the smallest Akaike Information Criterion (AIC) value, is then selected as the ARIMAX calendar effect model. Also, as a comparison, We tested the ARIMA model without involving calendar effects. The ARIMAX model developed through three stages of procedures: model identification, parameter estimation, and diagnostic checking [28].

b. Forecasting with ANFIS
  i. Determine the appropriate input variables ANFIS
     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, we will obtain all the input variables that will meet the LM test criteria. The optimization of variable input stops when the LM test value is not significant for input addition. At this stage, considering the principle of parsimony, ANFIS architecture is used with two membership functions and two rules. The steps for using the Lagrange Multiplier Test for determining input variables are as follows
     1. Choose the first input variable that has the largest R^2.
     2. Estimating parameters in the restricted ANFIS model with the output variable Z_t and the first input variable candidate
        $$Z_t = \overline{w_1}(p_{11}Z_t^1 + p_{10}) + \overline{w_2}(p_{21}Z_t^1 + p_{20}) + \varepsilon_t.$$
     3. Calculates the estimated residual (ε_t) the value of the restricted model
        $$\hat{\varepsilon}_t = Z_t - \overline{w_1}(p_{11}Z_t^1 + p_{10}) + \overline{w_2}(p_{21}Z_t^1 + p_{20}).$$
     4. Enter an additional candidate variable input that is the candidate input variable with the next largest R^2 value.
     5. Form an unrestricted ANFIS model to increase the input variables (input becomes 2) with the restricted model's residuals being input variables
        $$\hat{\varepsilon}_t = \overline{w_1}(p_{11}Z_t^1 + p_{12}Z_t^2 + p_{10}) + \overline{w_2}(p_{21}Z_t^1 + p_{22}Z_t^2 + p_{20}) + v_t.$$
        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.
     6. Determine the conclusions of the hypothesis test.
        Hypothesis testing the determination of the addition of one input variable using the LM test as follow.
        $$H_0: \theta_1 = \theta_2 = \cdots = \theta_m = 0$$
        If the calculated LM value = $nR^2_{\varepsilon_t} > \chi^2_{(\alpha, df)}$ then H_0 is rejected, which means that we can use potential additional input variables as an input. Furthermore, increasing the number of input variables continues to be done to determine all the inputs that enter the ANFIS architecture.
  ii. Determine the optimal number of membership functions (MFs).
      The LM test to determine the number of MFs (cluster) begins by forming the ANFIS model by entering all the input variables selected in the previous stage. Then increase the number of clusters starting from 2 number of MFs then calculating the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), values of the ANFIS architecture that was formed. Furthermore, increasing membership functions to the optimal number of membership functions provide the smallest RMSE and MAPE values. Finally, choose the optimal number of membership functions that give the smallest RMSE and MAPE values.

  iii. Forecasting
       The initial stage is done by determining the input and output variables. We are further dividing the data into training (in samples) and testing (out samples). This study used Gauss as a membership function and the number of membership functions that satisfy the LM test obtained in the previous step. This study uses three performance measures, namely RMSE, MAPE, and R^2, to evaluate forecasting accuracy. The architecture with the smallest RMSE and MAPE and shows a high R^2 value is the best ANFIS architecture.

3. Results and Discussion

A simulation study is first carried out before applying the ARIMAX-ANFIS based LM test to predict real data. We used simulation data to prove that the proposed hybrid ARIMAX-ANFIS based LM test

could handle the same pattern in the time series model that affected the holiday effect. We generated simulation data with 300 sample size. Meanwhile, the real data is the monthly foodstuff price index data obtained from bps.go.id from 2010 to 2019.
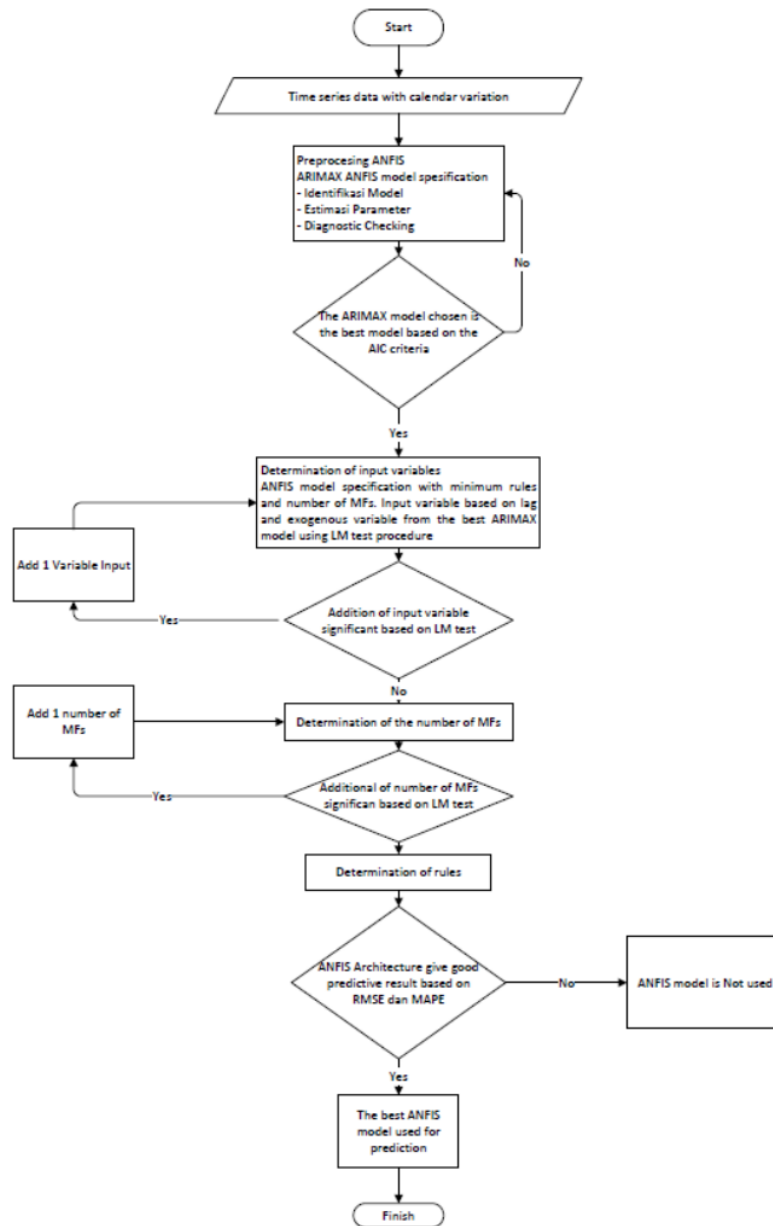

Figure 1. Flowchart of ARIMAX ANFIS Procedure Based on LM Test

Figure 2 shows a repeating increased pattern. The red line shows the period that the Eid holiday. There is always a spike at that period. This pattern indicates a calendar effect on the data. The simulation data was constructed following the ARIMAX model ([1,11,12], 1,0) with calendar dummy variables as exogenous variables. Meanwhile, the best model for the foodstuff price index data is ARIMAX ([1,11,12], 1,0) with a dummy variable calendar effect. ARIMAX shows that the variables that affect the two data are lag 1, lag 11, lag 12, and the calendar dummy for the month Eid al-Fitr occurs and the month after. Based on the ARIMAX results, the significant input variable candidates tested again using the LM test procedure to determine the ANFIS input variable. The LM test is used to re-select the input variables that actually affect the data.
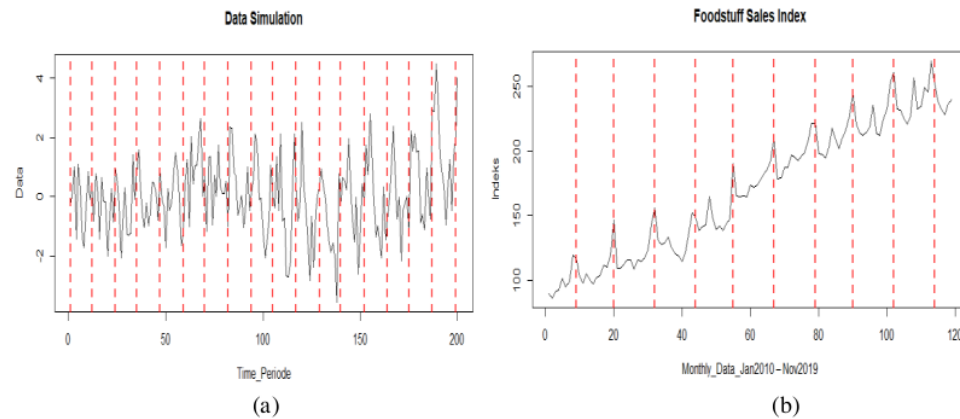

Figure 2. Time Series Plot with Calendar Effect (a) Simulation Data, (b) Foodstuff Price Index.

The model is formed by adding the input lag data and dummy variables one by one in sequence according to the value of R^2 of each variable. The variable with the largest R^2 value was tested first, followed by the next variable with the largest R^2 value below it. The input variables in both the simulation data and the foodstuff index data show that lag 12, lag 11, and lag 1, show a large R^2 value. The simulation data noted that lag 11, lag 12, then lag 1 and the dummy variable the month Eid al-Fitr occurred and the month after Eid al-Fitr have R^2 values of 0.197, 0.175, 0.124, 0.023, and 0.012, respectively. Meanwhile, in the foodstuff index data, the largest R^2 value is lag 12, followed by lag 11, lag 1, the dummy variable the month Eid al-Fitr occurred, and month after Eid al-Fitr with R^2 respectively is 0.9662, 0.9375, 0.9333, 0.0256, and 0.0.0000187. The process of determining the input variable is shown in table 1. Both data show that the LM test value exceeds the $\chi^2_{(\alpha,df)}$, so we reject H_0. We can see that all significant input candidates based on the ARIMAX model become the ANFIS input variables. Thus, there are five inputs for both simulation data and foodstuff index data.

In the simulation data, increasing the number of MFs shows that the greater the number of MFs made the RMSE smaller. The test results with the LM test show that we can use the addition of the number of MFs because the LM test value is more significant than $\chi^2_{(\alpha,df)}$. Because the number of parameters estimated is not recommended to be more than the observational data, 4 clusters are the maximum number of MFs that can be used. When we use 2 to 4 clusters in simulation data, the RMSE and MAPE values tended to decrease. Meanwhile, in the foodstuff price index, as the number of MFs increases, the RMSE and MAPE values are increased. We can see that 2 clusters give the smallest RMSE and MAPE values. Based on these results shown in table 2, it can be concluded that the number of MFs to be used is 4 for simulations data and 2 for the foodstuff price index.

Table 1. Determination of input variables

Simulation data			
Input Variable	**RMSE**	**LM Stat**	**Conclusion**
Lag 11	1.120	54.391	Reject H_0
Lag 11, Lag 12	1.061	28.585	Reject H_0
Lag 11, Lag 12, Lag 1	1,039	38.589	Reject H_0
Lag 11, Lag 12, Lag 1, D_{t+1}	1,031	42.231	Reject H_0
Lag 11, Lag 12, Lag 1, D_{t+1}, D_t	1,019	47.678	Reject H_0
Foodstuff Price Index			
Input Variable	**RMSE**	**LM Stat**	**Conclusion**
Lag 12	9.185	91.789	Reject H_0
Lag 12, Lag 11	195.955	14.372	Reject H_0
Lag 12, Lag 11, Lag 1	195.925	26.242	Reject H_0
Lag 12, Lag 11, Lag 1, D_{t+1}	195.906	34.896	Reject H_0
Lag 12, Lag 11, Lag 1, D_{t+1}, D_t	195.882	45.688	Reject H_0

Table 2. Determination of number of MFs (Clusters)

Input Variable	Num of MFs	RMSE	MAPE	LM Stat	Conclusion
		Simulation data			
Lag 11, Lag 12,	2	1.0197	15.041	92.296	Reject H_0
Lag 1, D_t, D_{t+1}	3	1.0053	14.959	7.710	Reject H_0
	4	0.9768	14.602	22.7151	Reject H_0
		Foodstuff Price Index			
Lag 12, Lag 11,	2	6.573	1.721	93.356	Reject H_0
Lag 1, D_t, D_{t+1}	3	195.978	66.332	10.203	Reject H_0

The ANFIS process begins by dividing the data into training and testing data. For simulation data, from 300 data, the training data is 213, while the rest is testing data. In the foodstuff price index, we select data from January 2010 to December 2015 as training data and data from January 2016 to November 2019 as testing data. Forecasting is done using an optimal architecture based on the previous LM test determination. As a limitation, the membership function used in ANFIS is Gaussian. This function was chosen because of its simple, with only two parameters (mean and variance) estimated. In a fuzzy system, each MF number is considered a rule. Therefore, the number of fuzzy rules is the same as the number of MFs.

Table 3. Training and testing ANFIS hybrid models

Model	Num of Input Var	Num of MFs	RMSE		MAPE		R^2	
			Training	Testing	Training	Testing	Training	Testing
			Simulation data					
ARIMA ANFIS	3	4	1.199	1.195	17.819	27.762	0.151	0.229
ARIMAX ANFIS	5	4	1.164	1.107	17.605	25.792	0.200	0.238
			Foodstuff Price Index					
ARIMA ANFIS	3	2	11.462	13.650	3.804	2.957	0.897	0.361
ARIMAX ANFIS	5	2	10.214	14.936	3.298	3.729	0.918	0.630

Table 3 shows the results of ANFIS training and testing using input and the optimal number of MFs. The optimal architecture for the simulation data is obtained using five input variables and four number of MFs. Meanwhile, the optimal foodstuff price index data with five input and two number of MFs. As a comparison, testing was also conducted for the model ARIMA ANFIS without considering the holiday

effect. We can see that the RMSE and MAPE for the ARIMAX ANFIS model give smaller values when compared to the optimal architecture of the ARIMA ANFIS. The simulation data shows that the MAPE value is high enough for the ANFIS training and testing process. However, [29] categorized this MAPE value in the criteria of reasonable forecasting. However, it can be argued that the calendar effect can reduce the predicted error value. In this study, it has been shown that the effects of holidays can significantly influence data patterns. When there are calendar variations, it is necessary to have an initial approach to the data to identify the model and consider the calendar variations' influence to get the best model.

4. Conclusion

This paper examines the problem of holiday effects that can occur in the monthly time series. ARIMAX is proposed to be used at the data preprocessing stage to determine the ANFIS input variable. Based on the test results, we can consider the LM test as an alternative way to define the input variable and the number of MFs in ANFIS. We can evaluate variables that significantly affect and eliminate variables that do not influence the data to obtain the optimal number of input variables. Likewise, in determining the number of MFs, the optimal number of MFs can be selected so that the ANFIS architecture gives the best results. The LM test is useful for obtaining an optimal yet simple ANFIS architecture with fewer parameters to be estimated. The results show that time series data analysis by including holiday effects will give better results than without considering calendar effect variables in the calculation. It is indicated by the smaller RMSE and MAPE values of the ARIMAX ANFIS model. The ARIMAX ANFIS method proposed can be an alternative way to predict time series data with calendar effects. Furthermore, it is also necessary to do ARIMAX ANFIS by forming a dummy variable that considers the proportion and number of days in a month. Furthermore, ARIMAX-ANFIS can be tested using several membership functions other than Gaussian.

References

Hybrid ARIMAX-ANFIS Based on LM Test For Prediction Of Time Series With Holiday Effect

banking system of Asia's next economic dragon? Evidence from rolling DEA windows", Applied Economics, 2014
Publication

| 7 | digitalpress.ugm.ac.id
Internet Source | <1% |

| 8 | studentsrepo.um.edu.my
Internet Source | <1% |

| 9 | mafiadoc.com
Internet Source | <1% |

| 10 | vdoc.pub
Internet Source | <1% |

| 11 | Mingi Kim, Han-Saem Kim, Choong-Ki Chung. "A Three-Dimensional Geotechnical Spatial Modeling Method for Borehole Dataset Using Optimization of Geostatistical Approaches", KSCE Journal of Civil Engineering, 2020
Publication | <1% |

| 12 | Marieta Monica, Agus Suharsono. "Forecasting cash outflow and inflow in Jember with ARIMAX calendar variation effect", AIP Publishing, 2021
Publication | <1% |

Exclude quotes   Off   Exclude matches   Off

Hybrid ARIMAX-ANFIS Based on LM Test For Prediction Of Time Series With Holiday Effect

FINAL GRADE

/0

GENERAL COMMENTS

Instructor