Modeling Threshold Linear in Transfer Function to Overcome Non Normality of the Errors

Heni Kusdarwati* and Samingun Handoyo1

1Department of Statistics, Faculty of Mathematics and Natural Science, Brawijaya University Malang

*Corresponding author: kusdarwati_h@yahoo.com

Abstract. The purposes of the paper are to apply a model of Double Seasonal Autoregressive Integrated Moving Average with transfer function (DSARIMAX) on the number of selecta tourist visitors in Batu city, which contain multiple seasonal elements and to add a threshold element to the DSARIMAX model. The DSARIMAX model produced is a feasible model but it has a low p-value because the assumption of normality of the errors is not fulfilled. Addition of the threshold to the DSARIMAX model makes the model feasible with high p-value and satisfies the normality assumption of the errors.

Keywords: double seasonal, modeling, threshold, time series, transfer function.

1. Introduction

Time series data are available in abundance in the real world. The needing of predicted values accurate generally uses a modeling time series data and cues the model for forecasting purposes. There have been many time series forecasting models that were developed by combining several methods including statistical methods, artificial intelligence and optimization techniques. Kusdarwati and Handoyo [1] hybridizing between discrete wavelet transform and Radial bases function, Handoyo and Marji [2] hybridizing between Fuzzy system and Ordinary Least Squared, and currently, Handoyo et al. [3] combines the fuzzy c-means and fuzzy system method. The applications of this sophisticated hybrid models have been able to produce predictions with high accuracy. However, the needed information for providing excellent service to people who visit tourism objects are not only based on accurate predictive values. There are several factors that must also be considered and known for their effects including the effect of school semester holiday, Idul Fitri holiday, Christmas and New Year holiday.

Recreational park is one of the attractions that will be crowded in the long holiday season. Data on the number of tourists visiting the Selekta Batu recreational park in East Java is influenced by long school holiday which are seen as seasonal influences with a 6-month period, a long holiday for Christmas and New Year's holidays has a 12-month period. These two holidays can be seen as double seasonal influences. The long holiday of Idul Fitri has period of 12 months less 11 days which can be seen as a holiday variation. The ARMAX model of holiday variations of the determinist trend by entering a double season, namely the length of school holidays and the Christmas and New Year holidays on the residuals was carried out by Janitra [4]. The result of the diagnostic testing of residuals model with the Q statistic was obtained a feasible model but had a low p-value. The sample autocorrelation plot (ACF) is still found to have significant lag autocorrelation values, so that the model is not feasible. In addition, the assumption of the normality of the model residuals cannot be fulfilled.
According to Cryer and Chan [5], if there is a significant sample autocorrelation (ACF), and the residuals do not meet the normality assumption, it can be overcome by a non-linear threshold model. In the article we build a linear threshold transfer function (ARMAX) model of holiday variation in deterministic trends with double seasonal residuals model. The resulted model is expected to be able to increase the p-value and be able to meet the assumption that residuals have normally distribution. Data set of the number of visitors visited to the Selecta Recreational Park in Batu of East Java were used as case studies.

2. Literature Review
The calendar variation model was developed according to the characteristics in the time series where the model is applied. The beginner was the model of Bell and Hillmer [6] which was a calendar variation model with the residual ARMA model. Suhartono et al. [7] developed a variation of the ARIMAX model calendar with the residual seasonal ARIMA model.

A calendar variation model according to Bell and Hillmer [6] is:

\[ X_t = D_t + N_t \quad \text{and} \quad D_t = \beta^{(i)} \frac{\sum h(i,t)}{\tau} \]

(1)

Where,
\[ D_t \]: effect of holiday variations at time of t
\[ N_t \]: Residuals from the calendar variation process formed in the ARIMA time series model
\[ \beta \]: parameters of the effect holiday variations
\[ \tau \]: long daily effects due to variations in the calendar of the dummy month variable
\[ i \]: duration of daily effects due to variations in calendar of dummy day variable
\[ t \]: time for holidays
\[ h(i,t) \]: dummy variable that represent daily effects of calendar variations

Suhartono et al. [7] added trend elements which were divided into deterministic trends, stochastic trends and seasonal elements in the residual model. The model of ARIMAX calendar variations with residual SARIMA model can be written as follows

a. Deterministic trend model

\[ X_t = \alpha + \beta_1 D_{i,1} + \ldots + \beta_n D_{i,n} + N_t \quad \text{and} \quad N_t = \frac{\theta_q(B)\Theta_q(B')}{\phi_p(B)\Phi_p(B')(1-B')^d} a_t \]

(2)

b. Stochastic trend model

\[ X_t = \beta_1 D_{i,1} + \ldots + \beta_n D_{i,n} + N_t \quad \text{and} \quad N_t = \frac{\theta_q(B)\Theta_q(B')}{\phi_p(B)\Phi_p(B')(1-B')^d (1-B')^d} a_t \]

(3)

where,
\[ t \]: variable that represent linear trend in the data
\[ \alpha \]: parameter of the trend variable
\[ D_t \]: dummy variable of holiday variation at t-time
\[ \beta \]: parameters of holiday variations
\[ N_t \]: residuals of the ARIMAX model with a calendar variations
d
\[ \phi \]: differencing order in the ARIMA mode
\[ D \]: differencing order on seasonal model

Based on the model obtained by Suhartono et al. [7] and the existence of a long holidays on Idul Fitri, semester holiday, Christmas, and new year, then the model from Lee needs to add one more seasonal element at the residual model. The model formed is a holiday variation ARIMAX model of
deterministic trend with residual double seasonal ARIMA (DSARIMA), it can be written with the following equation [4]:

\[ X_t = \alpha + \beta_1 D_{1,t} + \ldots + \beta_n D_{n,t} + N_t \]

\[ N_t = \frac{\theta_q (B) \Theta_1 (B^s) \Theta_2 (B^{s2})}{\phi_p (B) \Phi_1 (B^s) (1-B^s)^\delta_1 \Phi_2 (B^{s2}) (1-B^{s2})^{\delta_2}} a_t \]  \hspace{1cm} (4)

where,

\[ \frac{\Theta_1 (B^s)}{\Phi_1 (B^s) (1-B^s)^{\delta_1}} : \text{the first seasonal} \]

\[ \frac{\Theta_2 (B^{s2})}{\Phi_2 (B^{s2}) (1-B^{s2})^{\delta_2}} : \text{second seasonal} \]

The model in equation (4) has been applied to data on the number of visitors to Selekta Batu, East Java, Indonesia. The results of the diagnostic test obtained a feasible model through a joint residual autocorrelation test (Q statistic) but a low p-value value. The model is not feasible with an autocorrelation test (ACF residual) and does not meet the assumptions of normality.

According to [5], if there is a significant (irregular) residual autocorrelation (ACF) and residual values that do not meet the normality assumption, it can be overcome by a non-linear threshold model or mixture model. The TAR model is a generalization of the Autoregressive (AR) model that allows different areas for series, which depend on the value of the past (Fan and Qiwei, 2002). The TAR model is constructed from several linear time series models, namely the AR model in each area, which integrates into a model that has a threshold. The TAR (p, d, γ) model can be formulated as follows:

\[ X_t = \begin{cases} 
\alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p}, & X_{t-d} > \gamma \\
\beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_p X_{t-p}, & X_{t-d} \leq \gamma + a_t 
\end{cases} \]  \hspace{1cm} (5)

Where,

| d | : Delay |
|---|---|
| γ | : threshold (turning point) |
| a_t | : residual which is a random variable distributed N(0, \sigma^2) |
| \alpha_i | : autoregressive parameters if \( X_{t-d} > \gamma \) |
| B_i | : autoregressive parameters if \( X_{t-d} \leq \gamma \) |

3. Research Method
The analysis method uses 2 stages as follows:
Stage I: Modeling of linear transfer function (ARMAX) models of deterministic trend holiday variations with ARIMA (DSARIMA) double seasonal residuals,

1. Determine the dummy variable of the number of domestic tourists visiting Selecta Batu in the period January 2004 to December 2014 in the form of proportions as the effect of holiday variations where the month which does not have the effect of holiday variations equals zero.
2. Modeling the effects of holiday variations with multiple regression analysis: the number of domestic tourists visiting Selecta Batu as the output series, time t and some dummy variables in the form of proportions as effects of holiday variations as input.
3. Calculating the residual \( N_t \) of the deterministic trend model.
4. Testing stationarity of residuals against both variance and average. If the residuals are not stationary with respect to variance, the residuals are transformed using the Box-Cox
transformation. Testing of residual stationarity against the average is done using the Dickey Fuller test. If it is not stationary to the average, differencing needs to be done.

5. Model identification uses ACF and PACF plots on a residual series (N_t) that are stationary.

6. Estimating parameters of the linear transfer function (ARMAX) model of the deterministic trend holiday variation with the residual model of double seasonal ARIMA (DSARIMA), using the Maximum Likelihood estimation method and testing the significance of parameters.

7. Diagnostic testing of white noise and residual normality

Stage II: Adding threshold,

1. Determining delay by building a linear regression model between \( \nabla N_t^* \) and \( \nabla N_{t-1}^* \). The linear regression with minimum Mean Squared Error is chosen as the delay variable.

2. Setting threshold value (\( \Delta = 0 \))

3. Adding threshold in \( N_t \) which contain a double seasonal pattern to TDSARIMA

4. Estimating model parameters using the Maximum Likelihood Estimation method and testing the significance of the parameters.

5. Diagnostic testing of white noise and residual normality

4. Result and Discussion

The effect of linearly increasing deterministic trends and holiday variations is made in the model of linear deterministic trend holiday variations.

\[
X_t = 12884 + 331.31t - 97982D_{1,t-1} + 21346D_{2,t} + N_t
\]

where,

- \( t \): influence of time (deterministic)
- \( D_{1,t-1} \): dummy variation of the first vacation for 1 month before Idul Fitri holiday
- \( D_{2,t} \): dummy variation of the second vacation in the month when the Idul Fitri occurred
- \( N_t \): Residual at t-time

The residual (N_t) on the above model did not have Normal distribution so the assumption of residual normality was not be fulfilled.

4.1. Residuals modeling with double seasonal ARIMA (DSARIMAX)

The residual modeling to ARMAX with deterministic trend holiday variation and the adding double seasonal to be DSARIMAX model. The stationarity in variance is done based on the Box-Cox plot the residual needed transformation of \( \sqrt{N_t} = N_t^* \). Based on ACF seasonal differencing as \( \nabla N_t^* = (1 - B)^6 N_t^* \) are needed. Model identification is done by looking at the ACF and PACF plots.
Figure 1 explains ACF and PACF values of some lag from first differencing of residual ($N_t$).

Based on Figure 1 the values of autocorrelation and partial autocorrelation in lag 1 are not significant. It means that the residual($N_t$) is stationary with the ARIMA order of (0,0,0). Identification of the effects of double seasonal can be seen through two significant autocorrelation values, namely in the 4th lag, and 6th lag. In Figure 1.b. It can be seen that the significant partial autocorrelation values occur in the 4th lag, and the multiple of both 6th lag and 12th lag. Based on the results of the identification above, the tentative model is DSARIMA (0,0,0)(1,1,0)$_6$ (1,0,0)$_{12}$.

The estimating parameters of the model use maximum likelihood method resulted model of DSARIMA(0,0,0)(1,1,0)$_6$(1,0,0)$_{12}$. as follows:
\[ X_t = 22.3476 t + 11433.8 D_{1,t-1} + 6268.9 D_{2,t} \]
\[ + \frac{\sigma_t}{(1 + 0.8299 B^6)(1 + 0.3956 B^{12})(1 - B^6)} \]

where,
- \( t \): influence of time (deterministic)
- \( B^6 \): seasonal effects of semester school holidays
- \( B^{12} \): seasonal effects of the annual Christmas and New Year holidays

The residual diagnostic testing with the Ljung-Box Q statistic shows a feasible model but the residuals did not distribute normally. The P-value obtained on the Kolmogorov-Smirnov statistic is less than 0.01 that it is less than the significant level of 0.01. Based on the result, the residual has not met the normality assumption.

![Probability Plot of at Normal](image)

Figure 2. The ResidualsNormality plot of DSARIMA (0,0,0) (1,1,0) \(_6\) (1,0,0) \(_{12}\) model

In the data set of the number of tourists visiting Selecta have a double seasonal element that is not able to be captured in the model, so the resulted residuals were not Normally distributed. It can also be seen on the histogram presented in Figure 3. Based on the histogram in Figure 3, it is known that there are two peaks that indicates the presence of bimodal elements in the residuals so that the assumption of normality residuals is not fulfilled. There are two bars (left and right ends) which are located very far away, indicated by 2 outliers in the data. The condition is supported by data plots where there are two observation points away from linear lines, namely at the 104th observation, namely in February 2013 and 116th in February 2014. In February 2013 Selecta tourist visits increased because that month the Chinese New Year festival coincided, which can indirectly attract the number of tourists. On February 14, 2014 a natural disaster occurred, namely the eruption of Mount Kelud, so that the manager was forced to close all the rides because of being covered by volcanic ash which could endanger visitors. Selecta conditions that are not conducive cause a decrease in the number of visitors.
Figure 3. The Histogram of Residuals of DSARIMA (0,0,0) (1,1,0) (1,0,0)12 model

4.2. The adding threshold to model of DSARIMA (0,0,0) (1,1,0)6 (1,0,0)12
Threshold addition in linear transfer function r (ARMAX) with holiday variations of deterministic trends and two seasonal residual ARIMA (DSARIMA (0,0,0) (1,1,0)6 (1,0,0)12). Determine delay by doing auto-regression modeling $\nabla N_t^{*}$ against $\nabla N_{t-k}^{*}$. The selected delay is based on the highest determination coefficient value, namely $\nabla N_{t-6}^{*}$ which was shown in Table 1 as follows:

| No. | Auto regression | $\nabla N_t^{*}$ | $\nabla N_{t-k}^{*}$ | Std  | $R^2$ |
|-----|-----------------|-----------------|-----------------|------|------|
| 1   | $\nabla N_t^{*}$ | $\nabla N_{t-1}^{*}$ | 18512.1 | 0.1   |
| 2   | $\nabla N_t^{*}$ | $\nabla N_{t-6}^{*}$ | 14149.5 | 43.4  |
| 3   | $\nabla N_t^{*}$ | $\nabla N_{t-12}^{*}$ | 18558.4 | 7.1   |
| 4   | $\nabla N_t^{*}$ | $\nabla N_{t-12}^{*}$ | 18558.4 | 7.1   |

Determination of threshold based on selected delay $\Delta N_{t-6}^{*}$ is stationary so that the value of the selected threshold is zero, so

$$I_t = \begin{cases} 
1, & \nabla N_{t-6}^{*} > 0 \\
0, & \nabla N_{t-6}^{*} \leq 0 
\end{cases}$$

The estimating parameters of the holiday variation ARIMAX model with the double seasonal error threshold of model ARIMA (DSARIMA) (0,0,0) (1,1,0)6 (1,0,0)12.
Diagnostic testing of residual autocorrelation with a Ljung-Box Q statistic shows that the model is feasible. The residual model distributed normally with a p-value of 0.038 (Figure 4). The existence of two outliers that are not handled in the model also causes a low p-value value.

\[
Z_t = 499.99t - 34525.8D_{12, t-1} + 27824.0D_{2, t-1} + 0.5458Z_{t-6} + 0.2569Z_{t-12} + N_t \\
N_t = (1 - B^6)N_t^* = \begin{cases} 
27824.0 + 3812.4VN_{t-6}^* - 0.11VN_{t-12}^* - 0.53VN_{t-18}^* - VN_{t-24}^* > 0 \\
1818.3 - 0.5719VN_{t-6}^* - 0.63VN_{t-12}^* - 0.65VN_{t-18}^* - VN_{t-24}^* \leq 0
\end{cases} + a_t^* \\
a_t^* = \frac{a_t}{(1 - 0.55B^6)(1 - 0.26B^{12})}
\]

Diagnostic testing of residual autocorrelation with a Ljung-Box Q statistic shows that the model is feasible. The residual model distributed normally with a p-value of 0.038 (Figure 4). The existence of two outliers that are not handled in the model also causes a low p-value value.

**Figure 4.** Residual plot of model with threshold

The comparison of the feasibility model between the model linear transfer functions (ARMAX) holiday variation in deterministic that the residual trend of double seasonal ARIMA (DSARIMA) (0,0,0) (1,1,0) 6 (1,0,0)12, and that previous model added threshold is shown in Table 2. The adding threshold causes an increase in p-value both in the Ljung-Box test and in the Kolmogorov-Smirnov test.

**Table 2.** The p-value of Ljung-Box test and Kolmogorov-Smirnov test.

| Model | p-value of the Ljung-Box test | P-value of Kolmogorov-Smirnov test |
|-------|-------------------------------|-----------------------------------|
|       | Lag 6 | Lag 12 | Lag 18 | Lag 24 |                        |                        |
| (1)   | 0.1625 | 0.3950 | 0.1511 | 0.1193 |                        | <0.01                  |
| (2)   | 0.4744 | 0.6563 | 0.2648 | 0.1897 |                        | 0.038                  |

Based on the p-value in Table 2 shows that the addition of the threshold increases the p-value around 0.03. This refers to the results of the residual normality test fulfilled at the level of significance 0.01.
5. Conclusion
The addition of threshold in the linear transfer function (ARMAX) model of the holiday variation of deterministic trends with double seasonal residuals, ARIMA (DSARIMA) was able to overcome residual abnormalities and was able to increase the p-value in the L-ung Box test.

References
[1] Kusdarwati, H., and Handoyo, S. (2018). System for Prediction of Non Stationary Time Series based on the Wavelet Radial Bases Function Neural Network Model. *Int J Elec & Comp Eng.* Vol. 8 (No. 4): 2327-2337
[2] Handoyo, S., and Marji. (2018). The Fuzzy Inference System with Least Square Optimization for Time Series Forecasting. *Indonesian J Elec Eng & Comp Science.* Vol. 11 (No. 3): 1015-1026.
[3] Handoyo, S., Marji, Purwanto, I.N. and Jie, F. (2018). The Fuzzy Inference System with Rule Bases Generated by using the Fuzzy C-Means to Predict Regional Minimum Wage in Indonesia. *International J. of Opers. and Quant. Management.* Vol (IJOQM). Vol. 24 (No. 4): 277-292.
[4] Janitra, V. (2016). Pemodelan ARIMAX Variasi Liburan Trend Deterministik dengan Sisaan ARIMA Musiman Ganda. Skripsi Statistika FMIPA UB. Unpublished.
[5] Cryer, J.D., and Chan, K.S. (2008). *Time Series Analysis with Application in R.* Springer. Iowa.
[6] Bell, W.R., and Hillmer, S.C. (1983). Modeling Time Series With Calendar Variation. Journal of the American Statistical Association, Vol. 78 (No. 383): 526-534.
[7] Suhartono, Lee, M.H., and Hamzah, N.A. (2010). In Calendar Variation Model Based on ARIMAX for Forecasting Sales Data with Ramadhan Effect,” in Proceeding Regional Conference on Statistical Science. pp. 349-361.