Time Series Analysis and Forecasting of "Sarung Gajah Duduk" *Search Based on Google Trends Data Using Calendar Variation Model for Time Series Regression, Naïve Method and Machine Learning Methods*

Novita Dwi Utami (06211540000019)
Dr. Suhartono, S.Si., M.Sc., Dr.rer.pol. Dedy Dwi Prastyo, S.Si., M.Si., Dr. Dra. Kartika Fithriasari, M.Si.
Departemen Statistika, Fakultas Matematika, Komputasi, dan Sains Data,
Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
E-mail: gmnovitadwi@gmail.com

***Abstract*—The objective of this research is to develop calendar variation model based on time series regression method for forecasting time series data with Eid al-Fitr effect and compare them to other machine learning methods in term of accuracy and goodness of fit. The methods is applied for modelling and forecasting of real time series data, specially the percentage of search with keyword "Sarung Gajah Duduk" that recorded in google trends. The result show that time series regression with calendar variation produce smaller RMSE and MAPE compared to other mthods that have been used (naïve method, neural network and support vector regression).**

***Keywords*— Calendar Variation, Google Trends, Machine Learning, Sarung Gajah Duduk, Time Series Regression.**

I. INTRODUCTION

Forecasting is an activity of estimating an event that will occur in the future or a certain time. In most Islamic countries, many monthly economics and business time series data are subjected to two kinds of calendar effect, namely trading days and holiday effects (Suhartono, et. al, 2010). The effects of these calendar variations need special handling to do time series analysis. If the analysis is done directly without handling the calendar variation effects, then the forecast result can be biased. There are so many forecasting methods that have been used, either statistical method or machine learning method. The example for statistical method such as ARIMA, time series regression, decomposition, ETS, exponential smoothing. The examples for machine learning method are ANN, RNN, KNN, LSTM, SVR, RBF, GRNN and so on.

Makridakis, et. al, (2018) proved that statistical methods have better accuracy and the goodness of fit compared to machine learning method. Based on the research that they have done before, the overall sMAPE for statistical method is smaller than machine learning method. ETS method procude smallest sMAPE (7,12) and RBF method produce biggest sMAPE (15,79).

As the country with the biggest muslim populations in the world, Indonesia has a large consumption of muslim clothing, including sarong. One of the most popular sarong in Indonesia is "Sarung Gajah Duduk". Along with the development of technology, consumers often doing some searching about the brand through google search engine. The data used for this research is the recorded data in google trends from the percentage search of keyword "Sarung Gajah Duduk" in google. As the sarong is often used during Eid al-Fitr, so time series analysis will be carried out with calendar variations with regression time series and naïve method. Beside using statistical method, the analysis for this research will also use machine learning methods such as neural network and support vector regression (SVR) and then compare them in terms of accuray and the goodness of fit. Hopefully, the result of analysis in this research can provide information about the forecast or prediction of the percentage search with keyword "Sarung Gajah Duduk" so that the producer can find out the amount and interest demand for its product by the market or consumer.

II. LITERATURE REVIEW

A. Descriptive Statistics

Descriptive statistics is a way of collecting and presenting data so that it can provide useful information. Descriptive statistics are useful for providing initial information that has been collected and presented. Descriptive statistics can describe the characteristics of data in the form of graphs, data centering, data dissemination, and others (Walpole, 2007).

B. Time Series Analysis

Time series is a collection of observations of data sequentially in units of time. Time series method is a forecasting method that utilizes a pattern analysis of relationships between variables that will be estimated with time variables. The type and pattern of data is one of the important things that need to be considered in time series forecasting. In general there are four types of time series data patterns, namely horizontal, trend, seasonal, and cyclical (Hanke & Wichern, 2005).

C. Time Series Regression

Regression in time series has the same form as the general linear regression. Assuming output or dependent form y_t, untuk $t = 1,2,\ldots, n$, which is influenced by the possibility of input or independent data, where the input is a fix and known, this kind of relationship can be indicated by a linear regression. (Shumway & Stoffer, 2006). If y_t has a trend, the trend is used as input, that can be written as follows.

$$y_t = \beta_0 + \beta_1 t + a_t \qquad (2.1)$$

Where w_t is a residual, with IIDN assumptions with mean 0 and varian σ_w^2 . The seasonal data form can be written as follows.

$$y_t = \beta_0 + \beta_1 S_{1,t} + \beta_2 S_{2,t} + ... + \beta_s S_{s,t} + a_t \quad (2.2)$$

Where $S_{1,t}$, $S_{2,t}$, …., $S_{s,t}$ are dummy variables in seasonal form. For the example if the data is monthly, so there are 12 seasonal dummy variables, 1 dummy for 1 month. Linear regression model for data that have calendar variations can be written as follows.

$$yt = \beta_0 + \beta_1 V_{1,t} + \beta_2 V_{2,t} + ... + \beta_p V_{p,t} + a_t \quad (2.3)$$

Where $V_{p,t}$ is a dummy variabel for the effects of p-calendar variations.

D. Naïve Method

The Naive method is a forecasting technique that assumes the forecast of the next period data is the same as the data in the previous period, so the Naïve method formula is formulated as follows.

$$\hat{Y}_t = Y_{t-1} \quad (2.4)$$

For seasonal data can use *Naïve* with the formula as follows

$$\hat{Y}_{t+1} = Y_{t+1-s} \quad (2.5)$$

E. Neural Network

An artificial neural network is a "computational mechanism able to acquire, represent, and compute a mapping from one multivariate space of information to another, given a set of data representing that mapping". The back-propagation training algorithm is the most frequently used neural network method and is the method used in this study. The back-propagation training algorithm is trained using a set of examples of associated input and output values. The purpose of an artificial neural network is to build a model of the data-generating process, so that the network can generalize and predict outputs from inputs that it has not previously seen. The ANN is a black box model is a multi-layered neural network, which consists of an input layer, hidden layers, and an output layer. The hidden and output layer neurons process their inputs by multiplying each input by a corresponding weight, summing the product, and then processing the sum using a nonlinear transfer function to produce a result.

A neural network consists of a number of interconnected nodes. Each node is a simple processing element that responds to the weighted inputs it receives from other nodes (Kumar, et. al, 2018). The simplest networks contain no hidden layers and are equivalent to linear regressions. The coefficients attached to these predictors are called "weights". The forecasts are obtained by a linear combination of the inputs. The weights are selected in the neural network framework using a "learning algorithm" that minimises a "cost function" such as the MSE (Hyndman and Athanasopoulos, 2017). Once we add an intermediate layer with hidden neurons, the neural network becomes non-linear. The figure below shows the architecture of neural network with hidden layer.

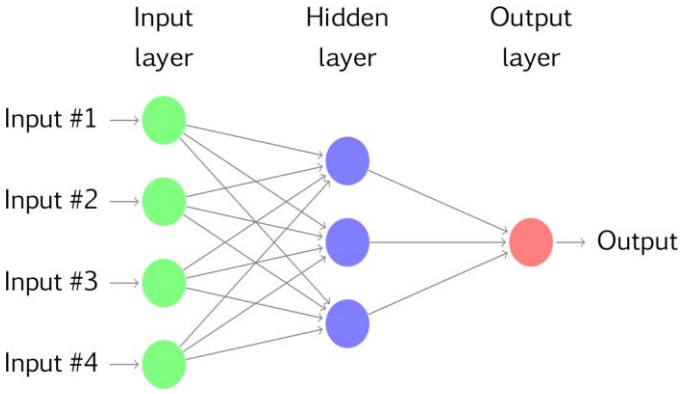

Figure 1. Architecture of Neural Network with Hidden Layer

F. Support Vector Regression (SVR)

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.

G. Sarung Gajah Duduk

Sarong is a wide piece of cloth sewn at both ends so that it is shaped like a pipe / tube. This is the basic meaning of the sarong that applies in Indonesia or places in the region. In terms of international clothing, sarong (sarong) means a wide piece of cloth that is used to be wrapped around the waist to cover the lower part of the body (waist down). One of the most popular sarong in Indonesia is "Sarung Gajah Duduk". "Sarung Gajah Duduk" produced by PT.Pismatex from Pekalongan, Central Java. "Sarung Gajah Duduk" received several Superbrand Awards. International Award, for the most popular and trusted brands in Indonesia. This award was obtained in 2004, 2008, and 2010. In 2010, "Sarung Gajah Duduk" also received the Topbrand award, the national award for the most trusted brand (Kasim & Hasanah, 2018).

III. RESEARCH METHODOLOGY

A. Data Source

This research used secondary data from website www.trends.google.com. The data is about the percentage of search for keyword "Sarung Gajah Duduk" during January 2011 to December 2017. The research data is divided into training and testing data. Training data is data from Januari 2011 to December 2016 while the testing data is data from January 2017 to December 2017

B. Reseacrh Variables

The research variable used in this research is the percentage of searches for keyword "Sarung Gajah Duduk" that are recorded in google trends.

Table 1. The Research Variables

Variabel	Keterangan
Y(t)	Percentage of Searches with keyword "Sarung Gajah Duduk" monthly

Table 2. Dummy Variables

	$M_i \begin{cases} 1 \\ 0 \end{cases}$	For Eid al-Fitr in the i-month Others
Calendar Variation	$W_{i,t} \begin{cases} 1 \\ 0 \end{cases}$	For Eid al-Fitr in the i-weeks and t-month, with i=1,2,3 Others
	$W_{i,t-1} \begin{cases} 1 \\ 0 \end{cases}$	For months before Eid al-Fitr in the i-weeks and t-month, with i=1,2,3 Others
	$W_{i,t+1} \begin{cases} 1 \\ 0 \end{cases}$	For months after Eid al-Fitr in the i-weeks and t-month, with i=1,2,3 Others

Table 3. Data Structure

Year	Month	Y(t)
2011	January	Y_1
2011	February	Y_2
.	.	.
.	.	.
.	.	.
2017	November	Y_{83}
2017	Desember	Y_{84}

H. Analysis Steps

The analysis step carried out in this research are as follows.

1. Collect the percentage of searches data of "Sarung Gajah Duduk" on the google trends.
2. Exploring data
3. Forecasting with time series regression method
4. Forecasting using the Naïve method
5. Forecasting using the Neural Network method
6. Forecasting using the Support Vector Regression (SVR) method
7. Compare the accuracy and goodness of the method that have been used.

IV. RESULT AND ANALYSIS

A. Descriptive Statistics

The following table shows the month of Eid al-Fitr for year 2011 until 2017.

Table 4. Weeks and Month of Eid al-Fitr

Year	**Eid al-Fitr**
2011	August, 3rd section week
2012	August, 2nd section week
2013	August, 1st section week
2014	July, 3rd section week
2015	July, 2nd section week
2016	July, 1st section week
2017	June, 3rd section week

From the table above, it can be seen that in year 2011 Eid al-Fitr occurs on the 3rd section week of August and so on for other months. There is a special pattern from Eid al-Fitr months. In every 3 years, Eid al-Fitr occurs in the same month and occurs in sequence in the third section week, second section week and then first section week. The graph below shows the percentage average search of keyword "Sarung Gajah Duduk" on google trend per year.

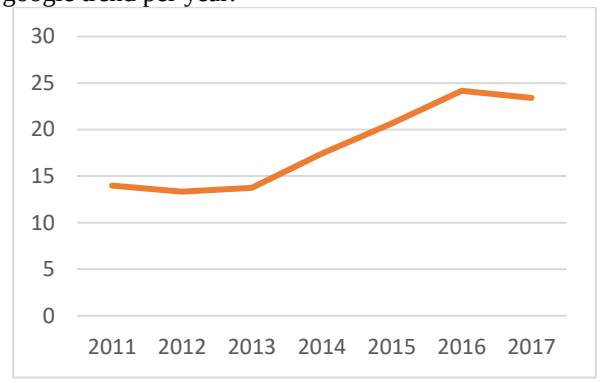

Figure 2. Line plot average Y(t) per year

The highest percentage average search of keyword "Sarung Gajah Duduk" occurs in 2016 and the lowest occurs in 2013. From the graph, it can be seen that there is a trend from the annual data. The graph below shows the time series plot for monthly data percentage search of keyword "Sarung Gajah Duduk" from Januari 2011 to December 2017.

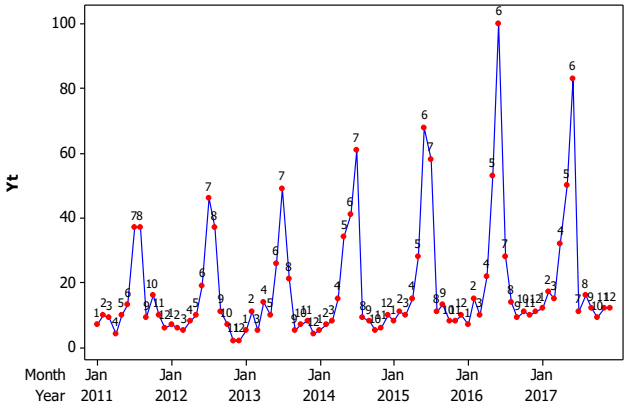

Figure 3. Time series plot of Y(t)

The time series plot shows that in 2011, the highest percentage search occurs in the 7th month and 8th month or in July and August. The highest percentage of all occurs in 6th month of 2016 (June 2016). The time series plot shows that the data contains trend and seasonal. Visually, the data contains multiplicative patterns. This time series plot will be a reference for analysis at the next stage and method.

B. *Time Series Regression*

The initial step which is based on the time series plot is used to identify calendar variation period affecting the data. The data are fitted with a linear regression model Y_t as the response (in percentage) ande the calendar variations effect as the predictors. for this, dummy variables are used for month and week where Eid al-Fitr occurs. The linear regression model is given by

$$Y_t = 0,17t + 1,22M_1 + 4,55M_2 + 2,21M_3 + 7,2M_4$$
$$+12M_5 + 8,61M_6 - 1,39M_7 - 3,06M_8 + 2,52M_9$$
$$+2,18M_{10} + 0,34M_{11} + 0,01M_{12} + 46,9W_{3(t)}$$
$$+31,2W_{3(t-1)} + 9,21W_{3(t-2)} + 43,3W_{2(t)} + 47,2W_{2(t-1)}$$
$$+7,17W_{2(t-2)} + 18,3W_{1(t)} + 62,6W_{1(t-1)} + 21,1W_{1(t-2)}$$

However, the ACf and PACF plot of error models, w_t in Figure 3 suggest that the model is inappropriate because the error is still not white noise.

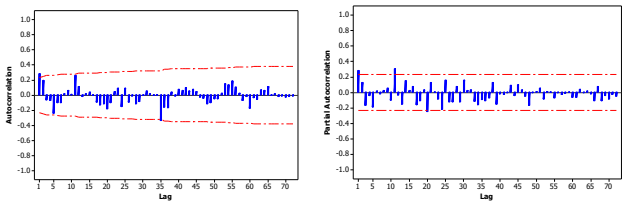

Figure 4. Time series plot of error Y(t)

From the ACF plot, the lags are significant in lag 1 and lag 35, so y_{t-1} and y_{t-35} are added into the model and the result is in the following fit.

$$Y_t = 0,173t - 1,27M_1 + 3,03M_2 + 1,22M_3 + 8,69M_4$$
$$+17,1M_5 + 14,6M_6 - 1,24M_7 + 2,60M_8 - 0,09M_9$$
$$-2,55M_{10} - 2,02M_{11} - 0,74M_{12} + 56,1W_{3(t)}$$
$$+20,3W_{3(t-1)} + 10,5W_{3(t-2)} + 51,3W_{2(t)} + 45,0W_{2(t-1)}$$
$$+2,22W_{2(t-2)} + 19,7W_{1(t)} + 75,4W_{1(t-1)} + 25,3W_{1(t-2)}$$
$$-0,0189lag_1 - 0,0071lag_{35}$$

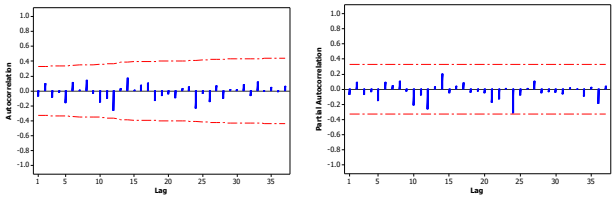

Figure 5. Time series plot of error Y(t) with lag

The error component from this model satisfies white noise assumption. But, check on the significance parameter shows that not all parameters are significant. The modeling process continues to eliminate insignificance parameters and re-estimate the model. The final calendar variation model based on time series regression with dummy variables is as follows:

$$Y_t = 0,0533t + 3,45M_4 + 12,5M_5 + 3,91M_6 + 22,6W_{3(t)}$$
$$-0,72W_{3(t-1)} + 1,36W_{3(t-2)} + 20,9W_{2(t-1)} + 1,17W_{1(t)} + 47,7W_{1(t-1)}$$
$$+14,0W_{1(t-2)} + 0,0824lag_1 + 0,827lag_{35}$$

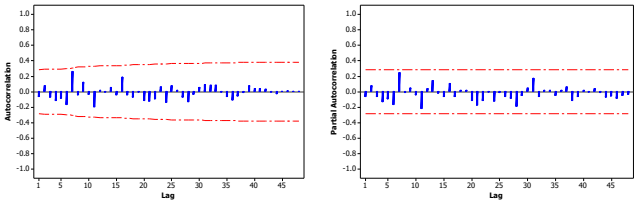

Figure 6. Time series plot of error Y(t) with significant model

From the ACF and PACF plots above, the error component from this model satisfies white noise assumption. So the last model is selected. From the last model, the forecast for the next 12 months are given below, with RMSE of out sample data is 3,767.

Table 5. Value of Forecast with TSR

	Forecast
January 2017	10,56
February 2017	11,55
March 2017	17,80
April 2017	38,21
May 2017	52,18
June 2017	85,18
July 2017	18,49
August 2017	11,79
Septsember 2017	9,77
October 2017	10,32
November 2017	13,43
December 2017	12,08

The plot below show time series plot for the next 12 months forecasting and out sample data (January – December 2017).

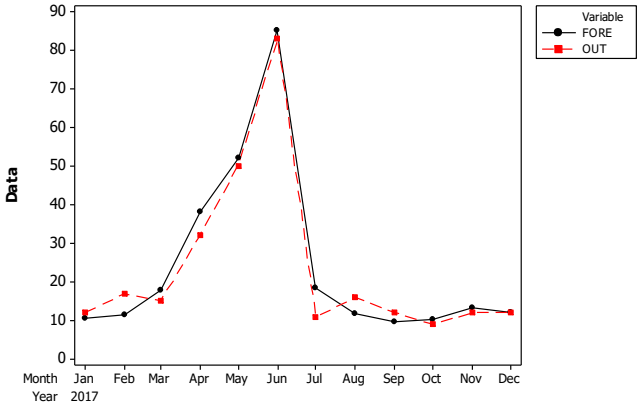

Figure 7. Time series plot of forecast TSR

C. *Naïve Method*

For naïve method, we devided the months into 2 parts, first is 3 months with calendar variation effects (the months where Eid al-Fitr occurs, the month before and the month after) and second is 9 months exclude the previous months. There is a seasonal and trend pattern indicated, where s=12 set for the seasonal index. The model for the other 9 months follows the N4 model as follows.

$$Y_{t+k} = Y_{t+k-s} + (Y_{t+k-s} - Y_{t+k-2s})$$

The naïve model for 3 months months with calendar variation effects (the months where Eid al-Fitr occurs, the month before and the month after) is given based on this table.

Year	Eid al-Fitr	Periode	3 Highest Months		
			I	II	III
2011	8/30/2011	W_3	8	7	6
2012	8/18/2012	W_2	7	8	6
2013	8/8/2013	W_1	7	6	8
2014	7/28/2014	W_3	7	6	5
2015	7/19/2015	W_2	6	7	5
2016	7/6/2016	W_1	6	5	7
2017	6/26/2017	W_3	6	5	4

2014, W_3 → 2015, W_2	2015, W_2 → 2016, W_1	2013, W_1 → 2014, W_3
$Y_t = Y_{t-s-1}$	$Y_t = Y_{t-s-2}$	$Y_t = Y_{t-s-1}$
$Y_t = Y_{t-s+1}$	$Y_t = Y_{t-s}$	$Y_t = Y_{t-s+1}$
$Y_t = Y_{t-s}$	$Y_t = Y_{t-s+2}$	$Y_t = Y_{t-s}$

From several model above, the forecast for the next 12 months are given below with RMSE of out sample data is 7,205.

Table 6. Value of Forecast with Naive

	Forecast
January 2017	6
February 2017	19
March 2017	10
April 2017	28
May 2017	53
June 2017	100
July 2017	-2
August 2017	17
September 2017	5
October 2017	14
November 2017	12
December 2017	12

The plot below show time series plot for the next 12 months forecasting and out sample data (January – December 2017).

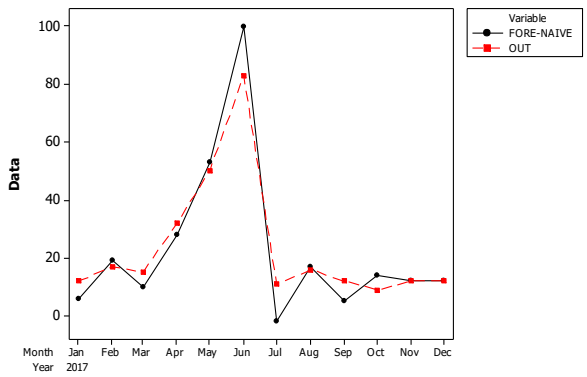

Figure 8. Time series plot of forecast Naïve

D. Neural Network (With Hidden Layer)

Time series analysis with neural network method in this research use 2 hidden layers and sigmoid activation function both in hidden layer and ouput layer. Before doing the analysis with neural network method, input must be determined first. The input is obtained from significant lags of data that has been stationary, both in mean and varians. The following steps show how to determine input for neural network method.

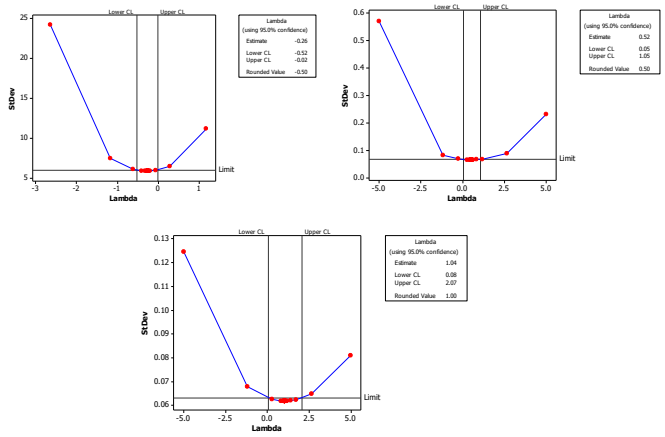

Figure 9. Box-Cox Transformation of Y(t)

From Figure 9 above, it is known that the data is stationary in variant after going through 2 transformation, first is sqrt transformation and the second is 1/sqrt transformation. After checking the variance stationarity, followed by checking the mean stationarity with ACF plot below

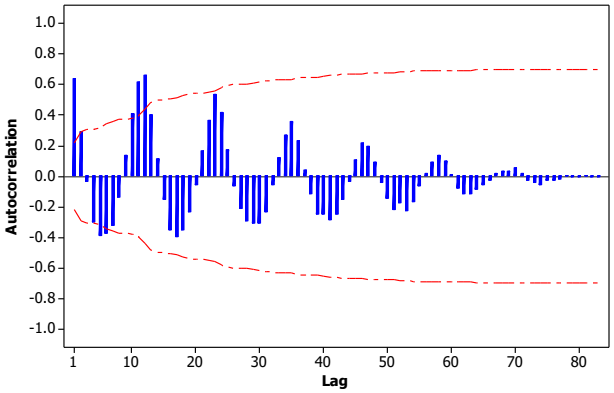

Figure 10. ACF of Y(t)

ACF plot indicate that the data is not stationary in mean because the ACF plot shows that there is a seasonal pattern. So, the data must be differenced in lag 12 (because the seasonal pattern). This plots below show the ACF and PACF plot from differenced data.

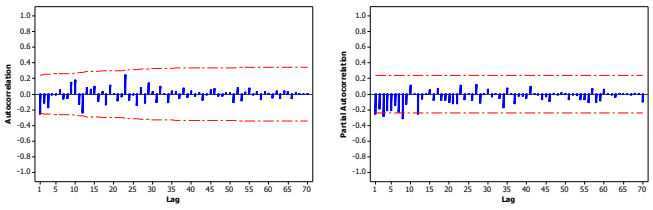

Figure 11. ACF of Y(t) after differencing

ACF plot from differenced data show that there is no seasonal pattern so it indicated that the data is stationary in mean. For the input, can be known from the significant lags in

PACF plot. The PACF plot shows that the lags are significant in lag 1, lag 3, lag 8 and lag 12. So the input for neural network method is lag 1, lag 3, lag 8 and lag 12. The figure below shows the architecture of neural network with 2 hidden layers.

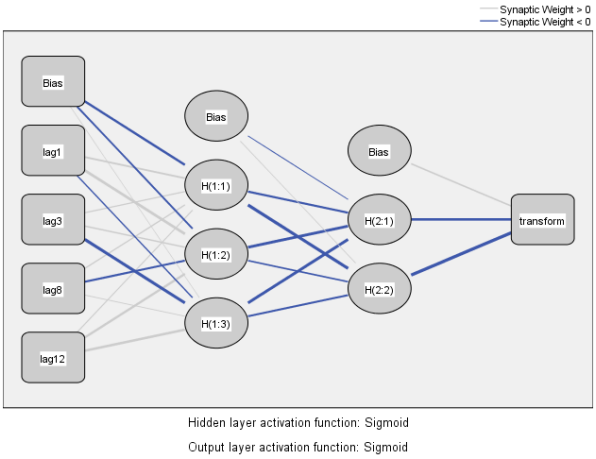

Figure 12. Architecture of NN with hidden layer

From the neural network architecture that has been built, can be determined the forecast for the next 12 months. The forecast are given in the table below where the RMSE of out sample data is 8,284..

Table 7. Value of Forecast with NN with hidden layer

	Forecast
January 2017	7,85
February 2017	16,15
March 2017	13,03
April 2017	35,60
May 2017	42,96
June 2017	57,28
July 2017	13,04
August 2017	9,88
September 2017	10,11
October 2017	13,23
November 2017	8,24
December 2017	11,48

The plot below show time series plot for the next 12 months forecasting and out sample data (January – December 2017)

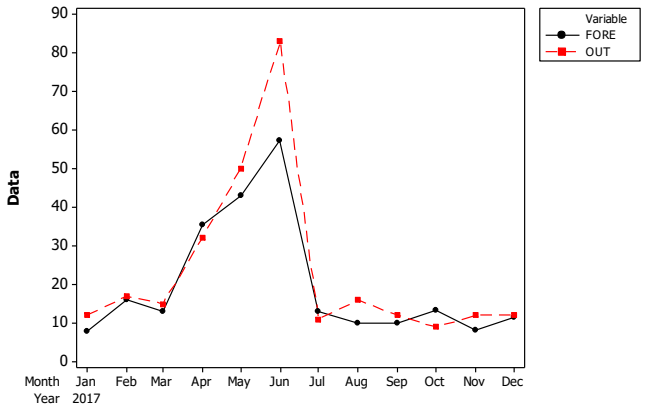

Figure 13. Time series plot of forecast NN with hidden layer

E. Neural Network (Without Hidden Layer)

The next method is Neural Network without hidden layer. Input fot this mthod is same as previous method (neural network with hidden layer). The input for this analysis are lag 1, lag 3, lag 8 and lag 12. The figure below shows the architecture of neural network without hidden layers.

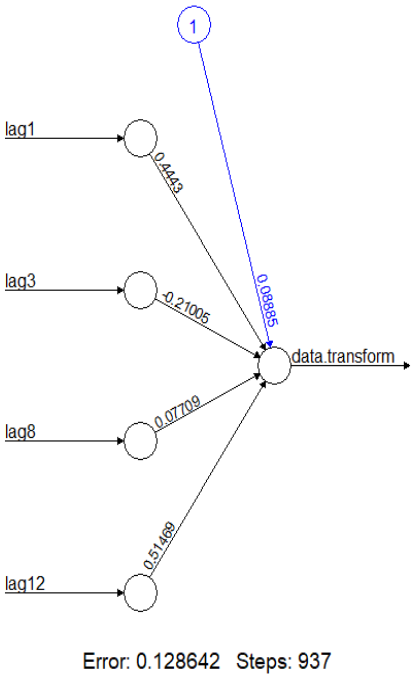

Figure 14. Architecture of NN without hidden layer

From the neural network architecture that has been built, can be determined the forecast for the next 12 months. The forecast are given in the table below where the RMSE of out sample data is 23,624..

Table 8. Value of Forecast with NN without hidden layer

	Forecast
January 2017	7,15
February 2017	9,07
March 2017	8,45
April 2017	9,63
May 2017	12,67
June 2017	15,73
July 2017	12,31
August 2017	7,16
September 2017	6,74
October 2017	7,94
November 2017	7,00
December 2017	8,02

The plot below show time series plot for the next 12 months forecasting and out sample data (January – December 2017)

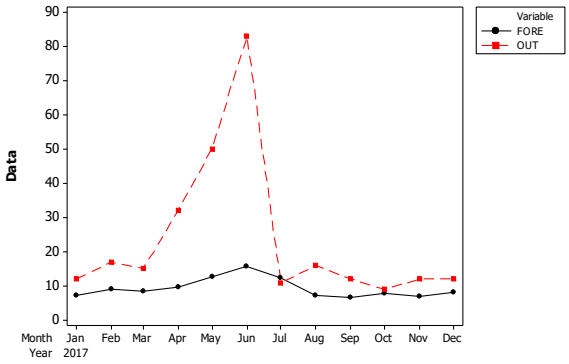

Figure 15. Time series plot of forecast NN without hidden layer

The time series plot above shows that there are many large differents of forecasting and out sample data. The line for forecasting doesn't fit the line for out sample data so it can be indicated that the method is less suitable for use in this research data.

F. Support Vector Regression (SVR)

The last method is Support Vector Regression (SVR). Input fot this mthod is same as previous method , the input for this analysis are lag 1, lag 3, lag 8 and lag 12. From Support Vector Regression (SVR) analysis, the forecast for the next 12 months are given below with RMSE of out sample data is 9,77.

Table 10. Value of Forecast with SVR

	Forecast
January 2017	9,91
February 2017	11,83
March 2017	12,13
April 2017	23,62
May 2017	50,48
June 2017	53,11
July 2017	20,76
August 2017	10,57
September 2017	11,26
October 2017	12,42
November 2017	9,84
December 2017	11,91

The plot below show time series plot for the next 12 months forecasting and out sample data (January – December 2017)

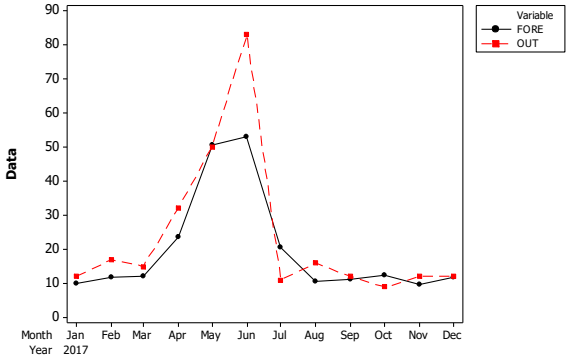

Figure 16. Time series plot of forecast SVR

G. Performance Evaluation for Each Method

The performance evaluation and accuracy comparison for each method in this research use RMSE and MAPE. The table below show value of RMSE and MAPE for each method that have been used.

Table 11. Accuracy Comparison

	RMSE	**MAPE**
Statistical Method		
Time Series Regression	3.767	18.544
Naïve	7.205	85.756
Machine Learning Method		
Neural Network (hidden layer)	8.284	26.916
Neural Network (no hidden layer)	23.624	127.766
Support Vector Regression (SVR)	9.767	28.027

From the RMSE and MAPE, we get the conclusion that time series regression with calendar variation has the smallest value of RMSE and MAPE among other methods. So for the data that is used for this research (Percentage of Searches with keyword "Sarung Gajah Duduk" that record in google trend data), the best method to forecast the percentage of searches for the next 12 month is time series regression with calendar variation. MAPE in range 10%-20% indicate that the forecasting ability of the model is good, so the forecasting ability of time series regression model with calendar variation is good because it produce MAPE 18,544%.

V. CONCLUSION AND FUTURE WORK

A. Conclusion

In general, time series data that is affected by calendar variation needs special treatment. The proposed models, i.e. time series regression models, yield better prediction compared to those of naïve method and machine learning methods (Neural Network and Support Vector Regression). It also proven that not all machine learning method produce better prediction (forecast) than statistical method. For "Sarung Gajah Duduk" data, time series regression with calendar variation provides a good model with RMSE 3,767 and MAPE 18,544.

B. Future work

This study proposed a procedure for building a calendar variation model based on time series regression method and comparing it to naïve method and some machine learning methods such as neural network and support vector regression (SVR). Future research is needed to validate this procedure to other real data sets.

REFERENCES

Hanke, J. E., & Wichern, D. W. (2005). *Business Forecasting Eight Edtion.* New Jersey: Pearson Prentice Hall.

Hyndman, R. J., & Athanasopoulos, G. (2017). *Forecasting: principles and practice, 2nd edition.* Australia: Monash University.

Kasim, H. R., and Hasanah, K. (2018). Strategi Penjualan Sarung Gajah Duduk Toko Megah Sutera di Pasar Sentral Makassar. *Journal of Business Administration.* Vol 1. No. 1

Kumar, D., Iakhwan, N., and Rawat, A. (2017). Study and Prediction of Landslide in Uttarkashi Uttarakhand, India Using GIS and ANN. *American Journal of Neural Network and Applications.* Vol 3. No. 6,pp 63-74

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concern and ways forward. *Journal PLOS ONE.* Vol 13. No. 3

Shumway, R.H., and Stoffer, D.S. (2006). *Time Series Analysis and Its Application with R Examples, second editio*n. Springer.

Suhartono, Lee, M. H., and Hamzah, N. A. (2010). Calendar variation model based on Time Series Regression for sales forecast : The Ramadhan effects. *Proceeding of the Regional Conference on Statistical Science.* pp 30-41

Walpole, R. (1995). *Pengantar Statistika Edisi ke-3 Terjemahan Bambang Sumantri.* Jakarta: Gramedia.