The Forecasting Technique Using SSA-SVM Applied to Foreign Tourist Arrivals to Bali

Yosep Oktavianus Sitohang*, Yudhie Andriyana, Anna Chadidjah
Department of Statistics, Padjadjaran University, Jawa Barat, Indonesia
*Corresponding authors, e-mail: yosep1707@gmail.com, y.andriyana@unpad.ac.id,
anna.chadidjah@unpad.ac.id

Abstract

In order to achieve a targeted number of foreign tourist arrivals set by the Indonesian government in 2017, we need to predict the number of foreign tourist arrivals. As a major tourist destination in Indonesia, Bali plays an important role in determining the target. According to the characteristic of the tourist arrivals data, one shows that we need a more flexible forecasting technique. In this case we propose to use a Support Vector Machine (SVM) technique. Furthermore, the effects of noise components have to be filtered. Singular Spectrum Analysis (SSA) plays an important role in filtering such noise. Therefore, the combination of these two methods (SSA-SVM) will be used to predict the number of foreign tourist arrivals to Bali in 2017. The performance of SSA-SVM is evaluated via simulation studies and applied to tourist arrivals data in Bali. As the results, SSA-SVM shows better performances compare to other methods.

Keywords: *Foreign tourist, Singular spectrum analysis, Support vector machine*

1. Introduction

Tourism is a prime sector in the Indonesian economy growth. In 2014 the tourism sector in Indonesia contributed 3.2 percent of the total national GDP and opened 3,326,000 jobs or 2.9 percent of total employment [1]. Therefore, in the National Medium Term Development Plan (RPJM) 2015-2019 the tourism sector becomes a priority sector [2]. Bali is a major tourist destination in Indonesia. It has a charm of natural beauty and cultural richness as the main attraction for tourists of both domestic and foreign. Especially, since 2015 many policies that support tourism in Bali, such as international events, visa-free policies, and the opening of new aviation routes from the most contributing countries of tourists in Bali such as Australia and China. Based on a data released by Statistics Indonesia (BPS), in 2016, the total number of foreign tourist arrivals to Bali was 44.88 percent [3]. In 2017, the Indonesian government set a target for the number of foreign tourists to be 15 million foreign tourists [4]. In order to achieve this target, a proper planning is required. Information on the estimated number of foreign tourists is needed in preparing the plan.

A common forecasting method used to predict the number of foreign tourist arrivals to Bali is Unweight Moving Average method [5]. This method can be said as a conventional time series forecasting technique because it is classified as a simple method. The performance of this method is also less favorable when applied to data containing trend, especially, a nonlinear trend [6,7]. One of forecasting methods that can overcome the limitations of Unweight Moving Average (MA) is Support Vector Machine (SVM). The SVM is one of the most recent machine learning methods introduced by Vapnik [8]. This method can be applied to a non stationary and non linear data [9]. This is because the SVM method applies the principle of Structural Risk Minimization (SRM) and uses ε-Insensitive Loss Functions.

The existence of an Incidental event or the change of government policy disturb on the data which can be categorized as noise. The noise tends to be irregular and unpredictable. This often makes the forecasting results of the model formed less accurate. Therefore, reducing the influence of noise components in building the model will certainly improve the accuracy of forecasting. In order to be able to implement it, the first necessary step is decomposing time series data into several components. This can be done using the Singular Spectrum Analysis (SSA) method [10]. This method has also been widely applied by some authors [11-13].

Received September 14, 2017; Revised January 13, 2018; Accepted July 2, 2018

In this research, we combine SSA with SVM (SSA-SVM) in predicting the number of foreign tourist arrivals to Bali in 2017. In order to see how far the SSA-SVM method can improve the accuracy of forecasting, the performances of the proposed method will be compared to MA, SSA and SVM techniques.

2. Research Method

2.1. The Foreign Tourist

A foreign tourist is everyone who is expected by a country outside his / her residence, which by one or several non-desired purposes he wishes to stay for no more than 12 (twelve) months [3]. This definition includes two categories of foreign guests, namely tourists and travelers. A tourist is a visitor stated above that stays at least twenty-four (24) hours, and will be no more than twelve (12) months in the place visited for the purpose of personal, business visits or professionals. A traveler stays less than twenty-four hours (including passenger cruise i.e. any visitor arriving in a country by boat or train, where they do not stay in the concerned country). The number of foreign tourist arrivals whose purpose in this study is the number of foreign tourists arriving through Ngurah Rai Airport Bali.

2.2. MA

The MA method is a simple forecasting method. This method predicts a value in a certain period by averaging a number of k values of the previous period [6]. Therefore, the accuracy of this method is determined by choosing the appropriate k value. Mathematically this method can be expressed as follows:

$$\hat{y}_{t+1} = \frac{y_t + y_{t-1} + ... + y_{t-k+1}}{k} \tag{1}$$

The y is an actual value, $\hat{y}$ is a forecasted value, t is time and k is an estimation period.

2.3. SVM

Consider $(\boldsymbol{x}_1, y_1),...,(\boldsymbol{x}_l, y_l)$, where $\boldsymbol{x} \in R^n$ is the input vector, and $y \in R$ is the corresponding values, and l is the amount of data. In the regression context, we consider the following model:

$$y = f(\boldsymbol{x}) + \varepsilon$$

where ε is a tolerated error and $f(\boldsymbol{x})$ is an unknown function, which can be formulated as follows:

$$f(\boldsymbol{x}) = \boldsymbol{\omega}^T \Phi(\boldsymbol{x}) + b \tag{2}$$

$\Phi(x)$ is a nonlinear function transforming $\boldsymbol{x}$ into a high dimensional space, $\boldsymbol{\omega}$ is a weight vector and b is a bias. The estimator of $f(\boldsymbol{x})$ is obtained by minimizing error risks ($R(f)$) using Structural Risk Minimization (SRM) as defined by:

$$R_{SRM}(f) = \tfrac{1}{2}\|\boldsymbol{\omega}\|^2 + \tfrac{1}{l}\sum_{i=1}^{l} L(y_i, f(\boldsymbol{x}_i)) \tag{3}$$

In equation 3, the first term $\tfrac{1}{l}\sum_{i=1}^{l} L(y_i, f(\boldsymbol{x}_i))$ is the empirical error and the second term $\tfrac{1}{2}\|\boldsymbol{\omega}\|^2$ is the regularization term. To obtain sparse solutions, Vapnik [8] introduce ε-Insensitive Loss Functions as follows:

$$L(y, f(\boldsymbol{x})) = \begin{cases} 0 & ; \text{if } |y - f(\boldsymbol{x})| \le \varepsilon \\ |y - f(\boldsymbol{x})| - \varepsilon & ; \text{otherwise} \end{cases} \tag{4}$$

The best estimator of $f(\boldsymbol{x})$ can be than obtained by minimizing following objective function:

$$\min_{\boldsymbol{\omega},b} \tfrac{1}{2}\|\boldsymbol{\omega}\|^2 + C\tfrac{1}{l}\sum_{i=1}^{l} L(y_i, f(\boldsymbol{x}_i)) \tag{5}$$

subject to $|y - f(\boldsymbol{x})| \le \varepsilon$
The final approximation function of $f(\boldsymbol{x})$ is:

$$f(\boldsymbol{x}) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{6}$$

subject to $0 < \alpha_i, \alpha_i^* < C$

$$\sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0$$

C is a pre-specified value in order to modulate the balance of empirical and regularization. α_i and α_i^* are the multiplier of Lagrange accordance with a support vector $\boldsymbol{x}_i$ and $K(\boldsymbol{x}', \boldsymbol{x})$ is defined as the kernel function. The kernel function employed in this study is Gaussian Radial Basis Function (RBF), because it has a better performance compared to the other kernel functions [14-15]. RBF is formulated by:

$$K(\boldsymbol{x}', \boldsymbol{x}) = \exp\left(\frac{\|\boldsymbol{x}' - \boldsymbol{x}\|^2}{2\sigma^2}\right) \tag{7}$$

There are three parameters (C, ε, σ) to be optimized. The optimization of those parameters uses a grid search technique. This technique is very powerful and able to improve the accuracy significantly [16]. The detailed information about SVM can be seen in Vapnik (1995) [8]. Applied to our time series data, we denote $\boldsymbol{x}$ in SVM by $x_t, x_{t-1}, x_{t-2}, ...$ and the output (y) replaced by x_{t+1}.

2.4. SSA

The SSA is divided into 4 processes namely embedding, singular value decomposition (SVD), grouping and diagonal averaging. The embedding process and SVD are known as decomposition stages as well as grouping and diagonal averaging reconstruction stages [10]. In the embedding process, a one-dimensional time series data, $x_1, ..., x_N$, will be replaced by the new multidimensional series, $\boldsymbol{x}_1, ..., \boldsymbol{x}_K$ where $\boldsymbol{x}_i = (x_i, ..., x_{i+L-1})^T \in R^L$, $1 \le i \le K$. Score L is in between $2 < L < N/2$ and L is commonly known as window length. The appropriate L value is gained from an optimization process and $K = N - L + 1$. New series data can be changed into matrix which is known as a trajectory matrix, as follows:

$$\boldsymbol{X} = [\boldsymbol{x}_1 : ... : \boldsymbol{x}_K] = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} x_1 & x_2 & \cdots & x_K \\ x_2 & x_3 & \cdots & x_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \cdots & x_N \end{pmatrix} \tag{8}$$

The trajectory matrix $\boldsymbol{X}$ will be changed into SVD. For instance, $\boldsymbol{S} = \boldsymbol{XX}^T$. From $\boldsymbol{S}$, we obtain eigenvalues $(\lambda_1,...,\lambda_L)$ with decreasing order of magnitude $\lambda_1 \geq ... \geq \lambda_L \geq 0$ and eigenvectors $(\boldsymbol{u}_1,...,\boldsymbol{u}_L)$ from each eigenvalues. Suppose, the rank of $\boldsymbol{X}$ is denoted by d, $d =$ max $\{i, \lambda_i > 0\}$ (note that in real-life series, usually have $d = L^*$ with $L^* =$ min $\{L, K\}$). Sequentially, it can be also gained $\boldsymbol{v}_i = {}^{\boldsymbol{X}^T \boldsymbol{u}_i} \big/ \sqrt{\lambda_i}$ for $i =$ 1,…, d, so that $\boldsymbol{X}$ can be formulated as follows:

$$\boldsymbol{X} = \boldsymbol{X}_1 + ... + \boldsymbol{X}_d = \sqrt{\lambda_1}\boldsymbol{u}_1\boldsymbol{v}_1^T + ... + \sqrt{\lambda_d}\boldsymbol{u}_d\boldsymbol{v}_d^T \tag{9}$$

$\boldsymbol{X}_i$ is called elementary matrix. The decomposed $\boldsymbol{X}$ matrix will be then grouped into m disjoint subsets $(I_1,...,I_m)$. If the $I = \{i_1,...,i_p\}$, so the results of $\boldsymbol{X}_I$ matrix corresponding to the group I, defined as $\boldsymbol{X}_I = \boldsymbol{X}_{i_1} + ... + \boldsymbol{X}_{i_p}$. The $\boldsymbol{X}_I$ matrix calculates the grouping of $I = I_1,...,I_m$ so that equation 8 can be elaborated as:

$$\boldsymbol{X} = \boldsymbol{X}_{I_1} + \boldsymbol{X}_{I_2} + ... + \boldsymbol{X}_{I_m} \tag{10}$$

The grouping process will be based on the eigenvectors ($\boldsymbol{u}$) plot and the cumulative ratio of eigenvalues (λ) [10]. Eigenvectors plot used to see the data characteristics of the elementary matrix and the cumulative ratio of eigenvalues used to see how much the contribution of the elementary matrix involved in the grouping process could explaining the condition of the trajectory matrix. The grouping results of (10) can be transformed to the new N time series data. This step aims to gain the single score of the data components obtained from grouping process. For instance equation 10 generated $\boldsymbol{Y}$ matrix whose size is LxK with element y_{ij}, $1 \leq i \leq L$ and $1 \leq j \leq K$. Let $y_{ij}^* = y_{ij}$ if $L < K$ and $y_{ij}^* = y_{ji}$ otherwise. Therefore the diagonal averaging enables us to replace $\boldsymbol{Y}$ into series $y_1,...,y_N$ which is formulated as follows:

$$y_k = \begin{cases} \dfrac{1}{k}\displaystyle\sum_{m=1}^{k} y_{m,k-m+1}^* & ;\text{ for } 1 \leq k < L^* \\ \dfrac{1}{L^*}\displaystyle\sum_{m=1}^{L^*} y_{m,k-m+1}^* & ;\text{ for } L^* \leq k < K^* \\ \dfrac{1}{N-k+1}\displaystyle\sum_{m=k-K^*+1}^{N-K^*+1} y_{m,k-m+1}^* & ;\text{ for } K^* < k \leq N \end{cases} \tag{11}$$

$L^* =$ min(L,K), $K^* =$ max (L,K), and $N = L + K - 1$.

2.5. SSA-SVM

The irregular and unpredictable noise components often lead to overfitting and underfitting. Therefore, referring to Wang *et al.* (2013), before analyzing the time series data using SVM, the noise component is firstly filtered using the SSA method [17] and then will be re-described according to the trend component and oscillation to create a better accuracy of forecasting. Hence, there will be three group of the data containing trend, oscillation and noise. The procedures of the SSA-SVM are illustrated by Figure 1.

2.6. Grid Search

The grid search method is one of the common methods to obtain the optimal (C, ε, σ) parameters in SVM and window length (L) in SSA method. We build some grid parameter points from a particular range [18]. We choose the optimal parameters corresponding to cross validation method for the time series data [19]. The best method is determined by smallest Mean Absolute Percentage Error (MAPE).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \qquad (12)$$

where y_i and $\hat{y}_i$ are respectively the actual and forecasted values and n is the number of data.

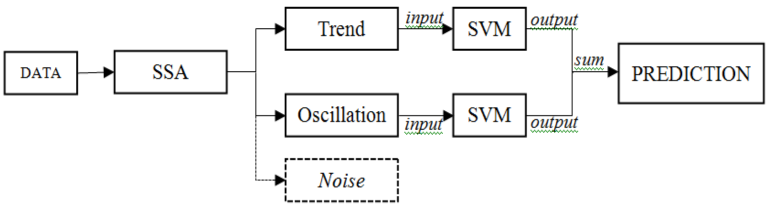


Figure 1. SSA-SVM Method

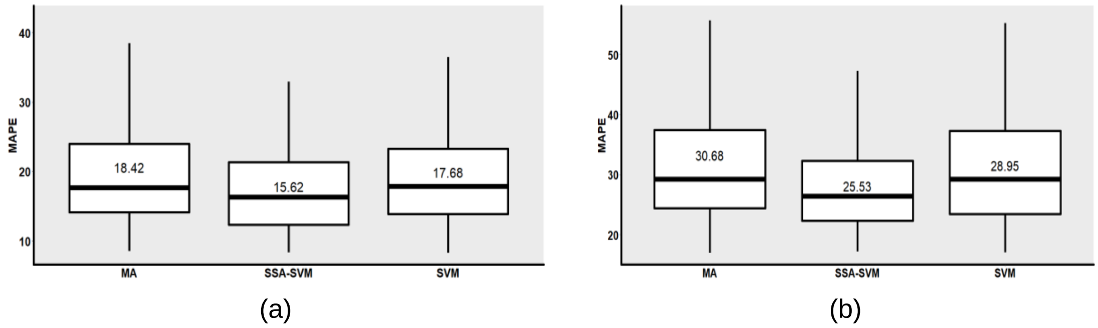
3. Results and Analysis

We analyze the data using some functions which available at **kernlab** and **Rssa** R packages. The performances of the proposed technique are evaluated via simulation studies. We also apply the propose technique to a real data application, in this case we apply to tourist arrivals data in Bali.

3.1. Simulation Study

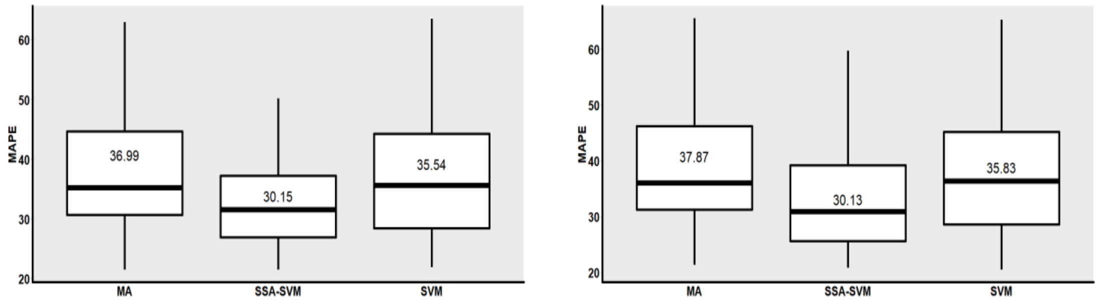
We simulate the generated model 200 times with the number observation (n) of each model is 74. The generated data is divided into training data (62 data) and testing data (12 data). The grid of required parameters are $C = \{10,20,30\}$, $\varepsilon = \{0,0.1,\ldots,0.7\}$, $\sigma = \{1,2,3,4\}$ and $L = \{2,3,\ldots,37\}$. Matrix $\boldsymbol{X}_1$ will be entered into the trend group, matrix $\boldsymbol{X}_2,\ldots,\boldsymbol{X}_5$ will be entered into the oscillation group and the rest will enter into the noise group.The predictions are applied to several periods of the data (3, 6, 9 and 12 data). It aims to predict how far the accuracy of forecasting from the two data types if the prediction range is longer.

The nonlinear data is generated from $y_t = 0.3 + y_{t-1} + 2\sin(2\pi t/16) + 2\cos(2\pi t/16) + u_t$ with $t = 1,2,\ldots,n$ and $u \sim N(0,1)$. Figure 2 shows that SSA-SVM method has a better performance inasmuch as the median and variety of MAPE resulted is smaller compared to the others method. The same is still be valid although the range of forecasting used is getting longer.



(a) (b)

Figure 2. Boxplot MAPE of Forecasting of Data for Prediction Periods of (a) 3, (b) 6



(c) (d)

Figure 2. Boxplot MAPE of Forecasting of Data for Prediction Periods of (a) 3, (b) 6, (c) 9 and (d) 12 Data

Based on the data simulation, SSA-SVM method has better performances compared to MA and SVM methods.

3.2. Real-Data Application

The data used in this study is foreign tourist arrivals data through Ngurah Rai (Bali) airport from January 2007 until December 2016. Figure 3 shows that the data has a positive trend and the right tail of the curve has a non-linear pattern. From the results of this initial identification, the MA method may imply a less accuracy.

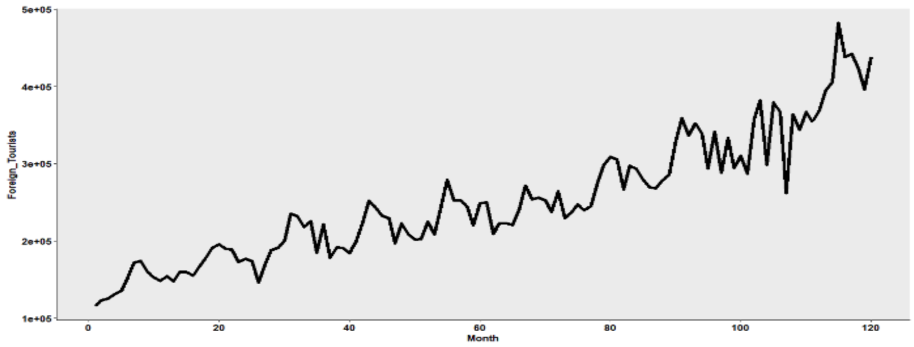
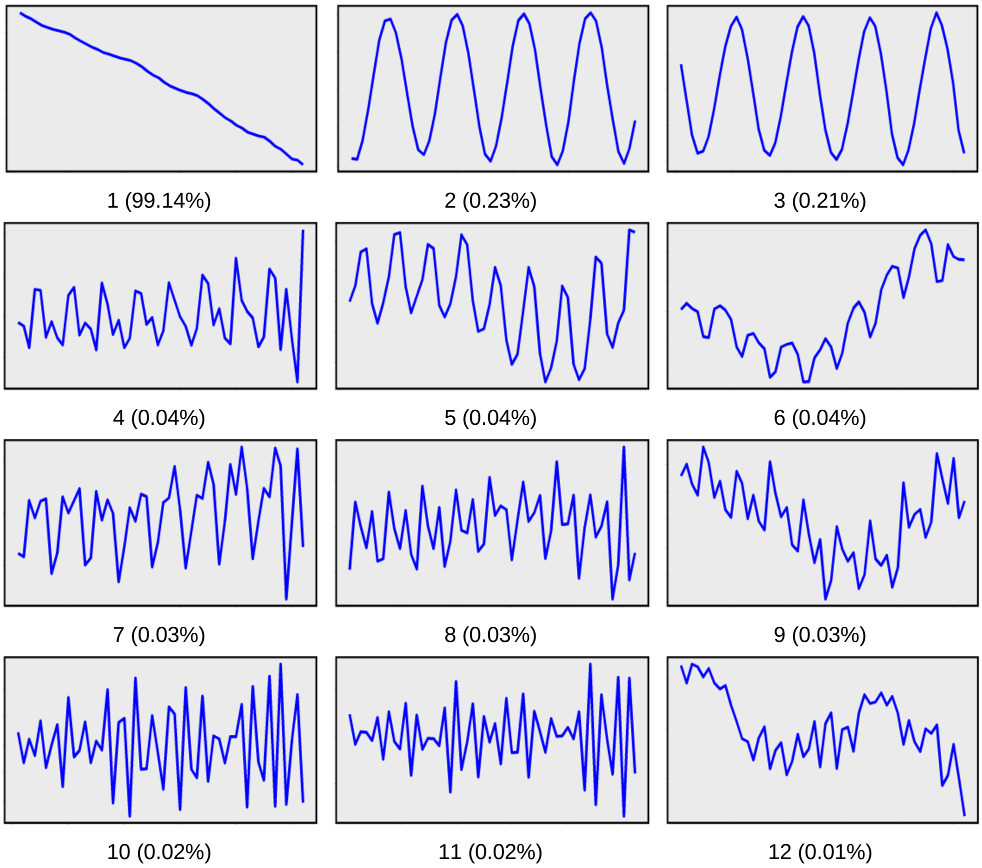


Figure 3. Plot of Data on the Number of Foreign Tourist Arrivals to Bali (Jan 2007-Dec 2016)

Before entering into the processing stage, the data are firstly divided into training data and testing data. The training data an of the number of foreign tourist arrivals from January 2007 to December 2015 (108 data) and data for the validation model (testing data) using data period January 2016 until December 2016 (12 data).

The grid search used with the same grid range as in the simulation data. As the results using SVM method we obtain $C=20$, $\varepsilon=0.7$ and $\sigma=1$. In the SSA-SVM method, window length obtained 52. Next the grouping process will be based from the eigenvector plot. This research does not show the whole plot of eigenvector but only 12 initial eigenvector plots, because it can already be represent the condition of the trajectory matrix. Based of Figure 4, matrix $\boldsymbol{X}_1$ has a trend pattern, so it be entered into the trend group. Matrix $\boldsymbol{X}_2,...,\boldsymbol{X}_5$ seen have a oscillation pattern, so it be entered into the oscillation group. And the rest will enter into the noise group. Furthermore, the value of cumulative ratio for 5 eigenvalue used, has reached 99.67 percent this indicates that using 5 elementary matrixs has been able to explain 99.67 percent condition of trajectory matrix. This already indicates that the grouping process performed is correct.



1 (99.14%) 2 (0.23%) 3 (0.21%)

4 (0.04%) 5 (0.04%) 6 (0.04%)

7 (0.03%) 8 (0.03%) 9 (0.03%)

10 (0.02%) 11 (0.02%) 12 (0.01%)

Figure 4. Plot 12 initial eigenvectors along with percentage of eigenvalue ratio

The optimal SVM parameters of this combined method for trend group are $C = 30$, $\varepsilon =$ 0.1 and $\sigma = 1$ and oscillation group are $C = 20$, $\varepsilon = 0.4$ and $\sigma = 3$. The MAPE for each forecasting methods can be seen in Table 2.

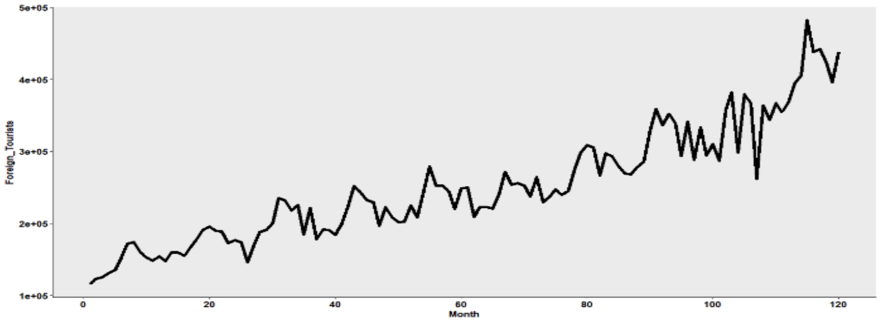


Figure 3. Plot of Data on The Number of Foreign Tourist Arrivals to Bali (Jan 2007 - Dec 2016)

Table 2. The MAPE Using MA, SVM and SSA-SVM Methods

Methods	MAPE (%)			
	3 months	6 months	9 months	12 months
MA	7.19	11.13	16.47	17.61
SVM	6.55	10.40	15.77	16.92
SSA-SVM	1.74	4.93	10.53	11.57

Table 2 shows that SSA-SVM method has the lowest MAPE in all prediction period. The SSA-SVM method has highly accurate forecasting for periods of 3 and 6 months and good forecasting for periods of 9 and 12 months [20]. Based on this, it can be said that the SSA-SVM method is the best method to predict the number of foreign tourist arrivals to Bali. The results of forecasting the number of foreign tourists, period January 2017-December 2017 using SSA-SVM method can be seen in Table 3.

Table 3. The Forecasting of Foreign Tourist Arrivals to Bali in 2017

Month	Foreign Tourists
January	443,342
February	454,101
March	454,883
April	452,444
May	460,026
June	436,759
July	447,737
August	463,621
September	429,521
October	451,844
November	461,504
December	433,435
Total	5,389,217

Based on Table 3, the number of foreign tourists who come to Bali in 2017 is estimated to reach 5.39 million tourists. We can also see that it has some fluctuations in every month where the highest number of arrivals is in August, but in overall it has a positive trend.

4. Conclusion

SSA-SVM method is the best method to forecast the number of foreign tourist arrivals to Bali. It combines the advantages of SSA that are able to decompose time series data to filtering out noise component and the ability of SVM method to handle nonlinear large variation of the data, it improves the forecasting accuracy.

Acknowledment

This research is supported by Statistics Indonesia (BPS) and master program of Applied Statistics, Padjadjaran University.

References

[1] World Travel & Tourism Council (WTTC). Travel & Tourism Economic Impact 2015 Indonesia. WTTC. 2015.
[2] Bappenas. Rencana Pembangunan Jangka Panjang Menengah Nasional 2015-2019. Jakarta: Bappenas. 2014.
[3] BPS. Statistik Kunjungan Wisatawan Mancanegara 2016. Jakarta: BPS. 2017.
[4] Republik Indonesia. Peraturan Presiden No.45 Tahun 2016 Tentang Rencana Kerja Pemerintah (RKP) Tahun 2017. Jakarta: Sekretariat Negara. 2016
[5] Kemenpar. Analisis Kunjungan Wisatawan Mancanegara pada Kawasan 3 Great Triwulan 1 2015. Jakarta: Kemenpar. 2015
[6] Lind DA, Marchal WG, and Wathen SA. Statistical Techniques in Business and Economics; Sixteenth Edition. New York: McGraw-Hill Education. 2002.
[7] Talluri KT, Ryzin GJ. The Theory and Practice of Revenue Management. New York: Springer. 2005.
[8] Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer. 1995.
[9] Gavrishchaka V and Banerjee S. Support Vector Machine as an Efficient Framework for Stock Market Volatility Forecasting. *Computational Management Science.* 2006; 39(2): 147-160.
[10] Golyandina N and Zhigljavsky A. Singular Spectrum Analysis for Time Series. New York: Springer. 2013.
[11] Sitohang YO and Darmawan G. (2017). The Accuracy Comparison between ARFIMA and Singular *Spectrum Analysis for Forecasting the Sales Volume of Motorcycle in Indonesia*. Proceedings of The

4th International Conference on Research, Implementation, and Education of Matheatics and Science (4th ICRIEM). Yogyakarta: AIP Conference Proceedings. 2017; 1868: 040011-1–040011-8.

[12] Hassani H et al. Forecasting U.S. Tourist Arrivals Using Optimal Singular Spectrum Analysis. *Journal of Tourism Management*. 2015; 46: 322-335.

[13] Unnikrishnan P and Jothiprakash V. Extraction of Nonlinear Rainfall Trends Using Singular Spectrum Analysis. *Journal of Hydrological Engineering*. 2015; 05015007(15): 1-15.

[14] Kim KJ. Financial Time Series Forecasting Using Support Vector Machines. *Neurocomputing*. 2003; 55: 307-319.

[15] Sotomayor A et al. Forecast Urban Air Pollution in Mexico City by Using Support Vector Machines: A Kernel Perfomance Approach. *International Journal of Intelligence Science*. 2013; 3(3): 126-135.

[16] Syarif I, Bennett AP and Wills G. SVM Parameter Optimization Using Grid Search and enetic Algorithm to Improve Classification Performance. *TELKOMNIKA (Telcommunication, Computing, Electronics and Control)*. 2016; 14(4): 1502-1509.

[17] Wang Y et al. Comparative Study of Monthly Inflow Prediction Methods for the Three Gorges Reservoir. *Journal of Stochastic Environmental Research and Risk Assessment*. 2013; 28(3): 555-570.

[18] Rao SS. Engineering Optimization: Theory and Practice. Fourth Edition. New Jersey (US): J Wiley. 2009.

[19] Hyndman RJ. Forecasting: Principles & Practice. Australia: University of Western Australia. 2014.

[20] Lewis CD. Industrial and Business Forecasting Methods. London: Butterworths. 1982.