Optimization of SARIMA Model
Using Genetic Algorithm Method in Forecasting
Singapore Tourist Arrivals to Malaysia

Mohd Zulariffin Md Maarof
Institute of Engineering Mathematics
Universiti Malaysia Perlis, Malaysia

Zuhaimy Ismail
Department of Mathematical Sciences, Faculty of Science
Universiti Teknologi Malaysia, Malaysia

Mohammad Fadzli
Institute of Engineering Mathematics
Universiti Malaysia Perlis, Malaysia

Copyright ©2014Mohd Zulariffin Md Maarof, Zuhaimy Ismail and Mohammad Fadzli. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract
This article proposes a numerical study of the genetic algorithm (GA) method in identifying the order and estimating the parameters of SARIMA model (GA-SARIMA). The primary goal of this study is to improve the traditional SARIMA model procedure for modeling Singapore tourist arrivals to Malaysia. Exponential smoothing and traditional SARIMA models were used as the benchmark model for model comparison. The findings of this study indicated that a combination of methods is more precise than a single process and GA can become an alternative in finding parameters and model orders of the SARIMA model.
Keywords: SARIMA, Genetic Algorithm, Tourist Arrival, Tourism Forecasting

1. Introduction

Tourism is regarded as one of the most rapidly growing global industries, and tourism forecasting is becoming an increasingly paramount activity in planning and managing the industry. The report released by the World Tourism Organization in 2014 exhibited an exponential growth in international tourist arrivals worldwide that grew at a rate of 5% during the first eight months of 2014, reaching nearly 781 million, compared to 36 million more than in the same period of 2013[1]. Consequently, both the academic interests and tourism literatures have grown, and parallel to this growth in the industry, many articles that model and forecast tourism flows between various countries were produced. These reports vary immensely in scope, modeling and forecasting techniques, and also in data types, lengths and frequencies [2], [3]. The tourism forecasts represent the most likely outcome for tourism given past trends, current information and the impact of policy and industrial changes.

Statistical and mathematical models could provide substantial contributions to the understanding and prediction of tourist arrivals and their trends of growth each year. Statistical tools such as time series analyses[4], [5], have been used by several authors to describe and forecast the number of tourists visiting in any country. Among these models, the seasonal autoregressive integrated moving average (SARIMA) model is useful in situations when the time series data exhibit seasonality-periodic fluctuations that recur with about the same intensity each year. This characteristic makes the SARIMA model adequate for studies concerning monthly tourist data, given that the number of tourist in a population tends to be subjected to seasonal variations, such that being maximum during vacation seasons and minimum during dry weather.

However, there are several shortcomings in the existing SARIMA model in terms of result optimization. The results derived from current SARIMA model can be optimized by combining artificial intelligence techniques such as Genetic Algorithms (GA)[6]. Genetic Algorithm is an adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. As such, they represent an intelligent exploitation of a random search used to solve optimization problems. Genetic Algorithms are by no means random; instead they exploit historical information to direct the search into the region of better performance within the search space. The basic techniques of GA were designed to simulate processes in natural systems necessary for evolution [7]

In this paper, a hybrid method was proposed by combining the SARIMA model which is a proven forecast methodology along with a state of the art, biologically inspired optimization technique called GA to increase precision in the prediction of tourist arrival. Overall, the objective of this study is to ameliorate
the conventional SARIMA model approach for modeling and observation of Singaporean tourist visits to Malaysia. In the experiment section, the proposed method was first applied on a given dataset; subsequently, the exponential smoothing (ES) model [8] and traditional SARIMA model were employed as benchmarking.

2. Data Set

The data used in this study is the Singapore tourist arrival into Malaysia provided by Malaysian Tourism Promotion Board. The data are monthly time series that cover from the year 2000 until 2013.

3. Methodology

In this study, three models were used which are the Box-Jenkins (BJ) family model (SARIMA), ES and GA. Previously, a large and growing body of literature had investigated on which forecast model is the most accurate model to forecast tourist arrivals in the tourism industry. The most frequently used time series in the tourism demand forecast literature is the BJ [9]. However, there is broad consensus that no one prediction model outperforms all the others in every occasion and that environment-specific conditions determine which method suits best an accurate forecasting task. This finding was supported with the results of the study conducted by Li et.al.[10], which showed that there is not a single forecast model that can outperform in any time series data.

This finding is also consistent with the findings of past studies as reviewed by Song and Li[9] whereby none of these models outperforms all of the others on all occasions. Thus, to improve the weakness of individual forecast model, combining the forecasts that are based on different methods or data has emerged as one of the most effective ways of improving forecasting performances [11]. Hence, the SARIMA model, GA and ES were applied in this study. The purpose is to verify whether the SARIMA model could be improved by combining it with GA method that acts as a tool for determining the parameter and identifying the SARIMA model order. The ES model is used as a comparison benchmark study. The three different models are described in the following subsections.

3.1 Seasonal ARIMA (SARIMA) Modeling

Box and Jenkins introduced SARIMA (1970) [12], and it is the most common model used in many applications and industries. By way of illustration, in tourism industry, Song and Li [9] reviewed the tourism forecast model since the year 2000 until 2007 and found that over two-thirds of the used prediction model was in fact, SARIMA model. SARIMA model is specified based on the standard Box-Jenkins method. This method incorporates seasonal autoregressive (SAR)
and seasonal moving average (SMA). If seasonal is excluded, ARIMA model will be applied.

In this study, monthly Singapore tourist arrivals to Malaysia have a seasonal time series feature. Hence, SARIMA with period $s(s=12)$ was used here due to the monthly data. The general formula for SARIMA model is as follows:

$$
\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D\hat{z}_t = \theta_q(B)\Theta_Q(B^s)a_t,
$$

where $B$ is a backward shift operator with $Bz_t = z_{t-1}$, and $Ba_t = a_{t-1}$. $z_t$ is a forecast value and $a_t$ is a white noise value at time period $t$.

\[\phi_p(B) = 1-\phi_1B-...-\phi_pB^p\] is a non-seasonal autoregressive of order $p$,

\[\Phi_P(B^s) = 1-\Phi_1B^s-\Phi_2B^{2s}-...-\Phi_PB^{Ps}\] is a seasonal autoregressive of order $P$,

while \[\theta_q(B) = 1-\theta_1B-...-\theta_qB^q\] is a non-seasonal moving average of order $q$ and last but not least, \[\Theta_Q(B^s) = 1-\Theta_1B^s-\Theta_2B^{2s}-...-\Theta_QB^{Qs}\] is a seasonal moving average of order $Q$.

### 3.2 Exponential Smoothing

During the past 60 years, much more information on exploring forecast models has become available in many industries. Brown [13] proposed the exponential smoothing method and it motivated Holt[14] to expand the work. Thus, it is known as the Brown’s simple exponential smoothing method.

Exponential smoothing methods performed quite well although the choice of exponential smoothing methods was very subjective and in some cases it was believed that the selected method is not the most appropriate. However, several studies to investigate exponential smoothing methods have been carried out towards modeling tourist arrivals to Australia [15] and also on seasonal effects of using Holt’s exponential smoothing model in investigating UK arrivals by air[16].

Hence in this study, the exponential smoothing method will be used as a benchmark to evaluate the forecast performance of the proposed model. Since the tourist arrival time series behavior has a seasonal pattern, winter’s exponential smoothing method is applied as follows:

The exponentially smoothed series or level estimate:

$$L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + T_{t-1})$$
The trend estimate:
\[ T_t = \beta (L_t - L_{t-1}) + (1 - \beta) T_{t-1} \]

The seasonality estimate:
\[ S_t = \gamma \frac{Y_t}{L_t} + (1 - \gamma) S_{t-s} \]

Forecast \( p \) periods into the future:
\[ Y_{t+p} = (L_t + pT_t)S_{t-s+p} \] (2)

where \( L_t \) is the new smoothed value or current level estimate, \( \alpha \) is the smoothing constant for the level, \( Y_t \) is the new observation or actual value in period \( t \), \( \beta \) is the smoothing constant for trend estimate, \( T_t \) is for trend estimate, \( \gamma \) is the smoothing constant for seasonality estimate, \( S_t \) is the seasonal estimate, \( p \) is the period to be forecast into the future, \( s \) is the length of seasonality and \( Y_{t+p} \) is for forecasting \( p \) periods into the future.

3.3 Genetic Algorithm and SARIMA Model

Genetic algorithms are acknowledged as right solutions for terrible problems. The process of implementing GA method in the model identification and estimation of SARIMA required fourteen steps. The steps are the initialization of total parameter maximum, estimation of the parameters, evaluation of fitness function, coding for genes, selection of operator, method for crossover, and mutation. The process can be summarized as follows:

i. Initialize the total number of maximum parameter of SARIMA model. The model used in this study is seasonal ARIMA\((p,d,q)(P,D,Q)^s\). The maximum order, \((p, q, P, Q)\) of SARIMA model is determined by analyzing the total number of significant lag on ACF and PACF using MATLAB programming. Then the exact order of SARIMA model is defined through GA method.

ii. Represent the chromosomes in four genes within the range of maximum order as determined in step one using an integer value where the total of all the parameters is equal to \((p+q+P+Q)\) and the range of the order is such that: \(0 < p < p_{\text{max}}\), \(0 < q < q_{\text{max}}\), \(0 < P < P_{\text{max}}\), and \(0 < Q < Q_{\text{max}}\).

iii. Generate the number of parameters randomly using GA based on the order needed in step two, in order to identify the best order based on fitness values.

iv. Determine the total size of the population and generation of chromosomes that are desired to be used.

v. Initialize the generation that have been formed.
vi. Calculate the values of fitness function of each chromosome.

vii. Select the best chromosome based on the fitness value that was calculated using the roulette wheel method. The best chromosomes will form a new population.

viii. Then, do the crossovers process by using single point crossover method. This process will produce quality chromosomes. Both of these methods are also applied in step six until a new population of high-quality chromosome is produced.

ix. Do the mutation process. At this stage, a new population by mutation process is formed.

x. Do the elitism process on each population by using type 1 and type 2 elitism. Then, a new fitness is recalculated until a new chromosome is formed.

xi. This process will stop when the maximum generation is reached.

xii. Determine the best model based on the high value of fitness function.

xiii. Determine the effectiveness and efficiency of the combined genetic operator based on the fastest convergence.

xiv. After identifying that the best order is achieved in step 3, randomly fine-tune the parameters of SARIMA model using GA and fix the identified order before repeating steps four to thirteen using the GA operator.

3.4 Forecast Evaluation

This study was designed to determine the best forecast model for modeling Singapore tourist arrivals to Malaysia. Therefore, to accomplish this aim, it is important to have forecast evaluation to measure the performance of a forecasting model. Thus, the mean square error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used in forecast evaluation. The MSE, MAE and MAPE formula are as follows:

\[ MSE = \frac{1}{N} \sum_{t=1}^{N} (z_t - \hat{z}_t)^2 \]

\[ MAE = \frac{1}{N} \sum_{t=1}^{N} |z_t - \hat{z}_t| \] (3)

\[ MAPE = \frac{1}{N} \sum_{t=1}^{N} \left| \frac{z_t - \hat{z}_t}{z_t} \right| \]

Where \( z_t \) is the actual value and \( \hat{z}_t \) is the forecast value at time \( t \), \( N \) is the size of the tourist arrival time series data. MSE is used in forecast evaluation to calculate the fitness value in genetic algorithm, SARIMA model and exponential model; while MAPE is represented as the performance of the forecast model for study comparison.
4. Results and Discussion

In order to prove the efficacy whether the proposed framework generates comparatively better results, a series of experiments have been designed. As seen in Figure 1, the Singapore tourist arrival monthly time series behavior has an upward trend, a little bit of seasonality and the data is not stationary. These patterns needed some modifications such that to be cleared in the preliminary analysis of the model. Thus, the data must follow certain criteria to find the best-fitted model.

![Figure 1](image-url)

**Figure 1:** Monthly Singapore tourist arrivals Malaysia for year 2000-2013.

When preliminary process had been done, the time series data will be converted to stationary in which the mean is zero and has constant variance. Then time series model which are the SARIMA model, exponential smoothing model and the proposed model, GA-SARIMA were applied as a mathematical model to suit the data in finding the best-fitted model. As we can see from Figure 2 and Table 1, the data is divided into two parts that are in-sample and out-sample. The in-sample is to measure the performance of the fitted model while out-sample is to measure the performance of the forecast model.

Surprisingly in Table 1, the traditional SARIMA model for in-sample was found to have the lowest mean absolute percentage error (MAPE) compared to the proposed model, GA-SARIMA and Exponential Smoothing model. This forecast accuracy shows that traditional SARIMA model is the best-fitted model for modeling monthly Singapore tourist arrivals to Malaysia. However, the observed difference between the SARIMA model and the GA-SARIMA model in this experiment was not significantly distinguished. The percentage difference between the two models is only 1.5352% of MAPE. The in-sample fitted models were illustrated as in Figure 3.
Figure 2: Comparison graph of monthly tourist arrival in-sample model for the year 2012.

Table 1: In-sample comparison between the proposed and benchmark models.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAPE</th>
<th>MAE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARIMA(0,1,1)(1,0,0)</td>
<td>8.3453%</td>
<td>88,166.74</td>
<td>12,511,783,497.8</td>
</tr>
<tr>
<td>GA-SARIMA(0,1,1)(1,0,0)</td>
<td>9.8805%</td>
<td>103,385.16</td>
<td>16,118,625,196.9</td>
</tr>
<tr>
<td>ES NA(0.718,0.258)</td>
<td>9.2170%</td>
<td>97,174.17</td>
<td>13,605,483,302.83</td>
</tr>
</tbody>
</table>

Figure 3: Comparison graph of monthly tourist arrival out-sample model for the year 2013.
Table 2 and Figure 3 clearly indicate that the proposed out-sample GA-SARIMA ranks itself as having the lowest MAPE position in comparison with the conventional SARIMA model because of the GA heuristic search behavior. It is interesting to note that GA simultaneously evaluates many points in the parameter space. Hence, it is more likely to converge toward the global solution. Thus, this highlights GA as a powerful mathematical technique to improve forecast accuracy.

<table>
<thead>
<tr>
<th>Model</th>
<th>Out Sample</th>
<th>MAPE</th>
<th>MAE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARIMA(0,1,1)(1,0,0)</td>
<td>Singapore</td>
<td>18.2539%</td>
<td>192,672.58</td>
<td>43,890,962,772.75</td>
</tr>
<tr>
<td>GA-SARIMA(0,1,1)(1,0,0)</td>
<td></td>
<td>12.8804%</td>
<td>147,042.35</td>
<td>37,841,574,638</td>
</tr>
<tr>
<td>ES NA(0.718,0.258)</td>
<td></td>
<td>20.7208%</td>
<td>220,656.75</td>
<td>52,086,146,961.75</td>
</tr>
</tbody>
</table>

GA-SARIMA for out-sample produced the lowest MAPE value which are 12.8804% compared to other models. These experimental simulation results indicate that GA can optimally improve the SARIMA model performance in terms of forecast accuracy for long-term forecast.

Table 3: Parameter values of GA-SARIMA model and SARIMA model.

<table>
<thead>
<tr>
<th></th>
<th>MA, φ</th>
<th>SAR, Φ</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARIMA(0,1,1)(1,0,0)</td>
<td>0.2656</td>
<td>0.4979</td>
</tr>
<tr>
<td>GA-SARIMA(0,1,1)(1,0,0)</td>
<td>-0.0007</td>
<td>-0.18097</td>
</tr>
</tbody>
</table>

As seen in Table 3, parameters for moving the average order 1 and seasonal autoregressive order 1, made from traditional methods have entirely different values when compared to parameters generated from proposed model, GA-SARIMA. A possible explanation for this difference in parameter values could be attributed to the different prediction error. Hence, it could be suggested that negative parameter values generated from the GA technique can improve the accuracy of forecast models.

5. Conclusions

The aim of the research work was to propose a combined methodology to prognosticate the international Singaporean tourist arrivals to Malaysia. In order to accomplish this goal, a combination of methodologies powered by SARIMA and GA was proposed. A time-series dataset was acquired from the Malaysian Tourism Promotion Board to perform the evaluation upon using the proposed algorithm. It
has been proven through experiments that the proposed method yields accurate results as compared with traditional SARIMA model. Additionally, the tests also indicated that to generate roughly accurate results, the amalgamation method of GA and SARIMA is a practical strategy in real-time environment with continuously changing data.

**Acknowledgements.** The authors gratefully acknowledge the Ministry of Education (MOE) Malaysia, SLAB/SLAI Universiti Malaysia Perlis (UniMAP) and Universiti Teknologi Malaysia (UTM) for their finance and administration assistance. This paper was supported by Fundamental Research Grant Scheme (FRGS) vote number 4F260 and FRGS/1/2013/ST06/UNIMAP/02/1 awarded by MOE.

**References**


Received: October 25, 2014, Published: November 28, 2014