Tourism forecasting using modified empirical mode decomposition and group method of data handling

N A Yahya ¹, R Samsudin ¹ and A Shabri ²

¹Faculty of Computing, Universiti Teknologi Malaysia, 81310, Johor, Malaysia
²Faculty of Science, Universiti Teknologi Malaysia, 81310, Johor, Malaysia

E-mail: ani@utm.my

Abstract. In this study, a hybrid model using modified Empirical Mode Decomposition (EMD) and Group Method of Data Handling (GMDH) model is proposed for tourism forecasting. This approach reconstructs intrinsic mode functions (IMFs) produced by EMD using trial and error method. The new component and the remaining IMFs is then predicted respectively using GMDH model. Finally, the forecasted results for each component are aggregated to construct an ensemble forecast. The data used in this experiment are monthly time series data of tourist arrivals from China, Thailand and India to Malaysia from year 2000 to 2016. The performance of the model is evaluated using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) where conventional GMDH model and EMD-GMDH model are used as benchmark models. Empirical results proved that the proposed model performed better forecasts than the benchmarked models.

1. Introduction
Tourism industry has contributed a lot to the social and economic development of Malaysia. Among its notable contribution includes assisting in escalating the rate of employment as well as increasing the gross domestic product (GDP) worldwide. Due to its importance, the requirement of having a systematic strategy is essential to cater for the increasing number of upcoming tourists. In such case, predictions of future events play a vital role in supporting the decisions and critical planning of the Malaysian tourism sector.

In previous literatures, various forecasting models have been utilized in tourism demand forecasting such as autoregressive moving-average (ARIMA), Exponential smoothing, Genetic Algorithm (GA) and Artificial Neural Network (ANN) [1]. Recently, artificial intelligence (AI) models have been applied in tourism forecasting to facilitate the limitations of conventional statistical forecasting models. In this research, the AI model of interest is the Group Method of Data Handling (GMDH) model, which is a subset of ANN model. This data-driven model is designed based on the survival-of-the-fittest concept whereby useful elements will be selected and underperforming elements will be discarded [2]. This model is ideal in modeling complex system without the modeler having to gain specific knowledge on the system [3]. Furthermore, GMDH model also has the ability to model non-linear data, which is important due to the fact that most real world data contains non-linear characteristics. Nevertheless, GMDH model does have its limitations such as its tendency in producing overly complex polynomials [3]. Hence, some previous researches tried to overcome this problem by hybridizing the GMDH model with other models [4].
Real world raw data often contains noise. Hence, pre-processing the data is necessary before applying the data to forecasting models to obtain more accurate results. The empirical mode decomposition (EMD) technique which was proposed by [5] has the capability of reducing the complexity of any time series by disintegrating it into smaller and finite components called Intrinsic Mode Functions (IMFs) and a residue. Among its other advantages is its application to both linear and non-linear data. Previous empirical results showed that hybridizing EMD with forecasting models yield positive results when applied on various time series [6, 7, 9]. Nevertheless, only several applications of EMD has been done in the area of tourism demand [6, 8]. Similarly, both researches proved that using EMD as a pre-processing tool improves the forecasting accuracy compared to models which did not apply EMD. However, the limitation of EMD lies in its inability to take into considerations the characteristic of the data during decomposition process.

For this research, a modified EMD-GMDH is proposed. The aim of the proposed model is to improve the existing decomposition method of EMD by reconstructing the decomposed components. Detailed explanation on the proposed model is explained in the following section.

2. Methodology

2.1. Group Method of Data Handling Model

GMDH is a heuristic self-organizing method used to construct extremely high order regression type polynomial. The conventional GMDH method is based on the fundamental assumption that the data used can be modeled by using Kolmogorov-Gabor polynomial as shown in the equation below:

\[ y = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m} \sum_{j=1}^{m} a_{ij} x_i x_j + \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} a_{ijk} x_i x_j x_k \ldots \]  

(1)

Based on the equation above, \( x \) represents the inputs to the system, \( y \) is the single output variable, \( a \) are the coefficients, while \( m \) represents the number of input variables. The design procedure for GMDH is described as follows:

**Step 1:** The input variables are defined where \( X = \{x_1, x_2, \ldots x_m\} \) and \( m \) is the total number of inputs. The dataset is divided into two parts; training and testing sets. The training set is used to build the GMDH structure while the testing set is used to evaluate the forecasting model.

**Step 2:** In constructing the GMDH structure, two inputs are combined at a time using the training data. The number of combinations for each layer is calculated using \( L=m(m-1)/2 \). In conventional GMDH, the model is developed using the polynomial:

\[ z = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2 \]  

(2)

Where the coefficients, \( a \) are obtained using regression.

**Step 3:** Least effective variables are eliminated by replacing the \( X = \{x_1, x_2, \ldots x_m\} \) column which is the old variable with the new variable, \( Z = \{z_1, z_2, \ldots z_m\} \).

**Step 4:** Repeat steps 2 to 3. The stopping criterion is done by checking whether the present layer performs better than the preceding layer. When the performance of the last layer starts deteriorating, the iterative computation is halted and the structure has been completed. Otherwise, the model will run recursively.

2.2. Empirical Mode Decomposition

EMD is a method applied in Hilbert-Huang Transformation (HHT) to process non-linear and non-stationary. Generally, EMD reduce the size of original data by decomposing it into smaller components which is referred to as intrinsic mode function (IMF) and a residue. In EMD preprocessing, the components must satisfy the following conditions in order to generate meaningful IMFs:

1. The sum of maxima and minima must be zero at all points.
2. The average of the envelopes defined by local maxima and minima must be zero at all points.

The algorithm for this technique is explained as follows:
Step 1: Each local maxima and minima for data series $y(t)$ where ($t=1,2,…,n$) is classified.

Step 2: Each local extrema is connected using spline interpolation to obtain the upper and lower envelopes, $y_u(t)$ and $y_l(t)$ respectively.

Step 3: The average of upper and lower envelopes is calculated using the formula

$$m(t) = \frac{[y_u(t) + y_l(t)]}{2}.$$ 

Step 4: Evaluate the difference between $y(t)$ and $m(t)$ where $z(t) = y(t) - m(t)$ to extract details.

Step 5: To make certain that $z(t)$ satisfies the two prerequisites of IMF, two details must be taken into considerations:

a. When $z(t)$ met the condition, an IMF is generated. $y(t)$ is then substituted with the residue as shown in the expression $r(t) = y(t) - z(t)$.

b. If $z(t)$ did not met the requirements, $y(t)$ will be replaced with $z(t)$ where $y(t) = z(t)$.

Step 6: The next IMF are extracted by applying the procedure above to residual term, $r_i(t) = y(t) - z_i(t)$ whereby $z_1(t)$ denotes the first IMF. The process is repeated until the final residue, $r_n(t)$ becomes a monotonic function and no more IMFs can be extracted from the original data. At the final stage, the data can be expressed by:

$$y(t) = \sum_{i=1}^{n} z_i(t) + r_n(t) \quad (3)$$

2.3. EMD-GMDH Model

The EMD-GMDH model is a combination of EMD and GMDH model. The process is explained as follows:

Step 1: The original time series data is fed to EMD to decompose the data into $k$-IMFs and a residue, $r_n(t)$.

Step 2: Use the GMDH model to forecast each component respectively.

Step 3: Aggregate the forecasting results of all the components to form the final forecasting result.

The diagram of EMD-GMDH model is illustrated as shown in figure 1 below:

![Diagram of EMD-GMDH model](image-url)

Figure 1. Diagram of EMD-GMDH model.
2.4. Modified EMD-GMDH Model
The procedure of the proposed method is explained as follows:

Step 1: Apply EMD technique on the time series data to decompose data into smaller components.

Step 2: Create a new component by reconstructing two components together. Since this method is done through trial-and-error, the components are added sequentially starting from lower frequency components to higher frequency components.

Step 3: Forecast the remaining IMFs and the new component respectively using GMDH model.

Step 4: Aggregate the forecasting results to produce the final forecast.

Step 5: Evaluate and save the results.

Step 6: Repeat steps 2 to 5 until no more new components can be created. The best result will be selected.

3. Data
The data used in this research are monthly tourism arrivals from China, Thailand and India to Malaysia which was obtained from Malaysian’s tourism authorities. The data are taken from January 2000 until March 2016, a total of 195 observations. In this study, the division percentage of data is 90% training and 10% testing.

4. Performance Criteria
To evaluate the performance of the models, root mean squared error (RMSE) and mean absolute percentage error (MAPE) are applied in this research. These criteria are among the frequently used performance measure [10]. The equations for both RMSE and MAPE can be defined as follows:

\[
RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2}
\]

\[
MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right|
\]

where \(y_t\) and \(\hat{y}_t\) is the actual and forecasted value respectively while \(n\) is the size of the sample. The performance of the model is judged based on the small values of RMSE and MAPE.

5. Results

5.1. Fitting EMD-GMDH Model to the Data
In this research, EMD is implemented on the time series data using EMD library in R software package. By using the EMD technique, the monthly tourist arrivals series are decomposed into several independent IMFs and a residue. All the IMFs generated are arranged from high frequency to low frequency components, and the final component represents the trend of the data as shown in figure 2. Each component is forecasted using GMDH model. By reducing the complexity of the original data, the forecasting accuracy could be improved through “divide and conquer” method.
5.2. Fitting Modified EMD-GMDH Model to the Data

In the first stage of the proposed modified EMD-GMDH model, EMD technique decomposes the tourism data into several IMFs and a residue. In the next stage, the components are reconstructed starting from the lowest frequency components, which is the residue and the last IMF generated. Then the remaining IMFs and the new component (reconstructed components) is forecasted individually using GMDH model. Finally, all the forecasts are aggregated to produce the final forecast which will be evaluated using RMSE and MAPE. The performance is saved to find the best result. The steps are repeated using other combinations for reconstructions. Due to this being a trial and error approach, the reconstruction is done in order of increase in frequency of components. To assess the performance of the proposed model, individual GMDH model and EMD-GMDH model are developed to serve as benchmark models. The performance results for tourist arrivals from China, Thailand and India are shown in Table 1.

Table 1. Performance of forecasting models for tourism arrivals data.

| Model               | China    | Thailand | India    |
|---------------------|----------|----------|----------|
|                     | RMSE     | MAPE     | RMSE     | MAPE     | RMSE     | MAPE     |
| GMDH                | 22641    | 9.2784   | 17924    | 10.811   | 6397.3   | 9.3992   |
| EMD-GMDH            | 14174    | 6.6339   | 9452.6   | 6.6279   | 5481.9   | 7.9871   |
| Modified EMD-GMDH   | 13982    | 6.6775   | 9237.2   | 6.2687   | 5371     | 7.9288   |

For China tourism arrivals data, EMD managed to improve the individual model’s accuracy by 37.4% for RMSE and 28.55% for MAPE. Meanwhile, for Thailand’s data, pre-processing using EMD improved the forecasting accuracy by 47.3% and 38.7% for RMSE and MAPE respectively. As for India’s tourism arrival data, the EMD technique reduced the value of RMSE and MAPE of individual model by 14.3% and 15% respectively. This indicates that by using EMD as a pre-processing tool, the accuracy of individual GMDH model can be significantly improved.

Referring to China’s tourism arrivals data in Table 1, the proposed model managed to further improve the accuracy of conventional pre-processing using EMD by 1.35% for its RMSE, however, the proposed method did not improve the MAPE of the data. Nevertheless, the proposed Modified
EMD-GMDH model managed to improve the accuracy of Thailand’s RMSE and MAPE by 2.28% and 5.42% respectively, while for India’s tourism data the improvements are by 2% and 0.73% respectively. Even though the improvements are quite subtle, it indicates that the proposed model has the potential to further improve the reliability of the conventional hybrid EMD and forecasting model.

6. Conclusions
The increasing importance of the tourism industry in Malaysia calls for extensive research in that area to anticipate the upcoming tourist arrivals. Hence, forecasting plays a vital role in assisting the decision making and planning. To further improve the current forecasting technique, this study proposed a modified EMD-GMDH model using monthly tourist arrivals from China, Thailand and India to Malaysia. Additionally, conventional GMDH and EMD-GMDH models are used as benchmarked models for performance comparison. Based on empirical results, the proposed model has the capability to slightly improve the result produced by conventional EMD-GMDH model for all the tourism data series. Provided with the right combination for reconstruction of components, it is believed that the proposed model can be a promising tool for forecasting tourism data.

Acknowledgments
The Malaysia Ministry of Higher Education has conferred Universiti Teknologi Malaysia the support for this research under the vote number R.J130000.7826.4F681.

References
[1] Cho V 2003 Tourism. Manage. 24 pp 323-30
Ma E, Liu Y, Li J and Chen S 2016 Tourism. Manage. Perspect. 17 pp 50-8
Palmer A, Montano J J and Sesè A 2006 Tourism. Manage. 27 pp 781-90
[2] Farlow S J 1981 Am. Stat. 35 pp 210-5
[3] Farlow S J 1984 Self-organizing methods in modeling: GMDH type algorithms (New York: Marcel Dekker)
[4] Samsudin R, Saad P and Shabri A 2010 Int. J. Res. Rev. Appl. Sci. 3 pp 274-9
[5] Huang N E, Shen Z, Long S R, Wu M C, Shih H H, Zheng Q, Yen N C, Tung C C and Liu H H 1998 Proc. R. Soc. A (London: The Royal Society) pp 903-95
[6] Chen C F, Lai M C and Yeh C C 2012 Knowl-Based. Syst. 26 pp 281-7
[7] Guo Z, Zhao W, Lu H and Wang J 2012 Renew. Energ. 37 pp 241-9
Lin C S, Chiu S H and Lin T Y 2012 Econ. Model. 29 pp 2583-90
Sang Y F, Wang Z and Liu C 2014 J. Hydrol. 510 pp 293-8
[8] Shabri A 2015 AIP Conf. Proc. vol 1643, ed M S Mohamad, W N Wan Yusoff et al (United States: American Institute of Physics) pp 192-200
[9] Wang X L and Li H 2010 Comput. Eng. Design. 31 pp 2303-7
Yu L, Wang S and Lai K K 2008 Energ. Econ. 30 2623-35
[10] Zhang G, Patuwo B E and Hu M Y 1998 Int. J. Forecasting. 14 pp 35-62