Prediction of Tourist Visit in Taman Negara Pahang, Malaysia using Regression Models

Sofianita Mutalib*, Athila Hasya Razali, Siti Nur Kamaliah Kamarudin, Shamimi A Halim, Shuzlina Abdul-Rahman
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA, 40450 Shah Alam,
Selangor, Malaysia

***Abstract*—Tourism is among the significant source of income to Malaysia and Taman Negara Pahang is one of the Malaysia's tourism spots and the heritage of Malaysia in achieving the Sustainable Development Goals (SDG). It has attracted many international and local tourists for its richness in flora and fauna. Currently, the information of tourists' visits is not properly analyzed. This study integrates the internal and public information to analyze the visits. The regression models used are multiple linear regression, support vector regression, and decision tree regression to predict the tourism demand for Taman Negara, Malaysia and the best model was deployed. Predictive analytics can support the decision-making process for tourism destinations management. When the management gets a head-up of the demand in the future, they can choose a strategic planning and be more aware about the factors influencing tourism demand, such as the tourists' web search engine behaviors for accommodation, facilities, and attractions. The factors affecting the tourism demand are determined as the first objective. The role of independent variable was set to the total number of visitors, subsequently being set as the target variable in the modeling process. A total of 30 models were generated by tuning the cross-validation parameters. This study concluded that the best model is the multiple linear regression due to lower root mean square error (RSME) value.**

Keywords*—Regression models; SDG; Taman Negara Pahang; tourist analytics

I. Introduction

Attractions of tourism destinations produce economic values as it impacts the number of tourist arrivals in Malaysia. As a result, tourism has become one of the highest revenue industries after automobiles and oil (REF). Today, one of the largest service sectors is the tourism industry where the industry became one of the highest in terms of revenue after automobiles and oil. This significant growth is a result of the efforts undertaken by the Ministry of Tourism where the planning and execution policy underlined by the government spearhead the success of the tourism industry. It is the long-term aim of the government to make Malaysia as one of the most popular tourism destinations. The success of the tourism industry is defined by the demand and supply which can be measured by tourists arrivals and receipts [1].

One of the famous tourism destinations is Taman Negara Malaysia, also known as Taman Negara National Park. This natural park protects a diverse flora and fauna, renowned for its nature trails, and adventure activities hence making it a valuable tourism source [2]. Encompassing an area of 4,343 km2, Taman Negara National Park straddles three states of Malaysia; Taman Negara Pahang, Taman Negara Kelantan, and Taman Negara Terengganu; in which Taman Negara Pahang takes up around 57% of the total national park area [3]. There are two different main entrances for Taman Negara Pahang namely, Kuala Tahan and Sungai Relau. There are many activities to do in Taman Negara such as jungle trekking, hiking, and fishing. Panoramic scene and captivating places such as waterfall cascades and canopy walkway attract many people to visit the Taman Negara National Park. Therefore, the service industry needs to be concerned about visitors' management of the tourism destination. In 2013, a research study was made about Ecotourism in Taman Negara National Park: the issues and challenges [4]. One of the issues that they found is the lack of visitor management especially on overcrowding problem and excess visitors during certain period of time. Moreover, based on the news [5], it was pointed out early on those statistics shows of slightly higher number of visitors to Taman Negara in January to February. The lack of proper service management on popular places at the park lead to overcrowding problem, mainly due to the inability to optimize staff's workload/working hours. In addition, overcrowding of tourists lead to the loss of authenticity and implies a significant risk to the destination's future attractiveness, especially towards vulnerable destinations such as the Taman Negara National Park.

The attractiveness of Taman Negara Malaysia as a tourism destination has been studied by Universiti Putra Malaysia [6]. The study evaluated that there are total of thirteen attractions in Taman Negara, namely oral history, local culture and lifestyle, flora, fauna, building architecture, nature trails, shopping opportunity, canopy walkway, caves, stream, fishing, mountain, and adventure activities [6]. The attractions of tourism destinations produce economic values as it also gives an impact to the number of tourist arrivals in Malaysia. These attractions have been characterized as demand structures. Thus, demand studies are needed for decision making support of tourism destination management.

Therefore, tourism demand forecasting is vital and will benefit the nation's tourism industry greatly. The study of regression techniques helps to forecast future and seasonal demands for tourism growth, management, and planning purposes. Regression analysis is a collection of statistical methods for estimating relationships between a dependent variable and one or more independent variables which can be used to determine the strength of a relationship between

*Corresponding Author.

variables and to predict how they will interact in the future [7]. This study was made to determine the factors affecting the tourism demand, to make a comparison study of regression techniques, and develop an analytical dashboard based on the best regression model that helps in forecasting tourism demands components, incorporating the applicable criteria in the following sub-sections.

A. Factors Affecting Tourism Visit

In demand studies, factors affecting tourism demand was determined as the independent variables that might affect the target variable, where the target variable is the total number of visitors. The comparison study was made to find the best model in predicting the target variable. Tourism industry has been a huge contribution factor to the economy, even though there exist many factors affecting tourism demand in Malaysia. Several studies have shown that the economic variables play an important role as the key economic factors. Based on [8] and [9], the key economic factors are exchange rate, income, consumer price index (CPI), and population of the country. The studies found a strong relationship between these key economic factors and the volume of tourists. However, the study also showed that there is a negative correlation between the exchange rate and the number of tourists, the higher the exchange rate, the lower the volume of tourists' arrivals. The exchange rate is related to the depreciation of Ringgit Malaysia (RM) that affects the cost of living in Malaysia.

As many researchers used economic factors as the contribution to the demands, some researchers [10-12] observed the tourist's web search behavior by using Google Trends data. Specific keywords were identified for web scraping related to tourism activities, such as "skiing", "skiing in sweden" or "sweden skiing" or "ski sweden" [10]. Meanwhile, [12] made use of these keywords: "destination", "destination + guide", "destination + travel guides", "destination + tickets", "destination + weather", depending on the tourism destination, strategy, ticket price, scenic spots, weather, and accommodation, among many other factors. At the end, 50 initial keywords related to the decision-making process were selected [11]. Due to the time gaps between web search activity and tourist arrivals, this new approach is truly relevant. Web-based data sources, such as search engine traffic, often have a natural relationship with tourism demand. Because of strong interest in certain tourism destinations, potential tourists for instance, browse websites extensively before visiting these destinations.

In other study, climate change is an additional factor to know the relationship with tourism demand. Research by [13] was conducted regarding the dimension of climate change in Malaysia based on tourists' perception. Generally, Malaysia has an equatorial climate. Extreme weather and seasonality are commonly related with climate change in Malaysia. Temperature, rainfall, and, to a smaller extent, wind are all examples of extreme weather factors. Seasonality, on the other hand, is always linked to the dry and wet (monsoon) seasons. The main variables for this factor are the average temperature and the average precipitation. The weather in Malaysia is hot and humid all year with Malaysia's average daily temperature ranges from 21°C to 32°C. Precipitation is the measure of the falling water from the sky which is the rainfall. The findings of the study revealed that tourists had sufficient knowledge of climate change, which influence their travel decisions [13].

B. Prediction in Tourist Domain

Many past research studies have been conducted in order to predict the tourism demand. A comparative study made by [14] to forecast the tourism demand in Turkey using data mining techniques based on regression modelling. The techniques used include multiple linear regression (MLR), multilayer perceptron (MLP) regression, and support vector regression (SVR). The author in [14] used monthly data points for their study, unlike previous studies which usually uses yearly or quarterly data. The author in [14] decided to choose these regression models because of the nonlinear mapping capabilities. In addition, the conventional methods like these models are more efficient to use for data that are likely to pattern the trends, seasonality, and cyclicality. SVR originated from a machine learning model hence a support-vector machine (SVM) can work for regression tasks and is suggested in order to forecast tourism demand. Unlike most conventional neural network model, SVR applies the theory of structural risk minimization which based on the idea of empirical risk minimization, to minimize the upper limit of the generalization error, instead of minimizing the error in training [15-17].

A research on tourists visit in the province of West Sumatra was done using MLR and Artificial Neural Network (ANN) [25] with inflation rates and Rupiah exchange rates. The results show an impressive accuracy within 96 to 99 percent. Other researchers from Indonesia made use of seven independent variables including the characteristics of foreign tourists (sex, age, occupation/profession, length of stay, nationality, purpose of visit, and accommodation) to identify the effect on total expenditure [26]. They also found that American and European tourists contributed to the largest average of the total expenditure for vacation purpose. Regression tree was used by [27] to segment the tourist length of stay in Barbados, with socio-demographic profile of the tourist, trip-related characteristics, distance, and economic conditions in the source country. Another study on tourism to model the revenue from international tourism using the foreign trade balance of the country shows the positive correlation in Azerbaijan example [28]. The result also showed that tourism will increase the country's foreign trade turnover. More advanced methods were explored in tourism domain in China [29] with social evaluation index as the attributes and hybrid methods of back propagation and fuzzy as the model.

Another angle on predicting tourism trend is by looking at the flow of tourists' movement in a specified area, as studied by [30-31]. Usage of user-generated content assisted in narrowing down specific criteria to forecast tourism demand apart from projecting possible point-of-interest for tourists [30]. By creating trajectory graphs on past data, the study by [30] yield a better result than traditional machine-learning based algorithms for forecasting the next tourist movement, which is useful for predicting tourism demand in certain areas. While the study by [31] focuses more on statistical-based techniques, they applied statistical method with BPNN model (SMBPNN) on variously collected data such as historical tourism flow, weather, and temperature. Their hybrid model which combines statistical technique with neural network

suggested an improvement of forecast as compared to stand-alone neural network models [31].

The methodology used by other researcher suggested using various techniques and at the same time, incorporating a multi-dimensional dataset. Hence for this study, an integrated dataset were created from various sources that will be described later in this paper. The suitability of the techniques with the available real-world dataset was also considered; hence this study will be focusing on application of regression techniques.

In the rest of the paper, we first provide some necessary background of proposed modelling methods in Section II, and the data source and evaluation are discussed in Section III. The experimental results are demonstrated and discussed in Section IV. Finally, the conclusion and future works are explained in Section V.

II. Modelling Methods for Regression

Regression analysis is used for predicting real values, for instance, to forecast the daily sales of the business which makes the number of sales as the target or dependent variable. To determine the relationship between the variables of interest, the data collected will be trained using a regression model. In order to determine the optimal model, Goodness-of-fit measurements, such as the square of the correlation coefficient (r^2 or squared correlation), was used to examine the scatter of data points around the fitted value. The number denoted the percentage of variance in one variable that can be explained by the other. The higher the r^2 score, the more precise the prediction. However, the number does not tell how precise the forecasts were in the dependent variable's units.

A. Multiple Linear Regression

Multiple Linear Regression (MLR) aims to model the linear relationship between the independent (explanatory) variables and dependent (response) variables, whereas simple linear regression only has a single input to estimate the value of the coefficients used in data representation. MLR model helps to predict an outcome based on multiple explanatory variables provided with details [18]. The representation of multiple linear regression will be like the following:

$$Y = a_0 + a_1x_1 + \ldots + a_nx_n \quad (1)$$

In which Y refers to the dependent variables, x_1, …, x_n represent independent variables, a_1, …, a_n as regression coefficients, and a_0 is y-intercept (a constant term).

By measuring slope and regression coefficients, it can be represented in the form of mathematical equations in multiple linear regression. Using the regression coefficient formula, the intensity and direction of the relationship between the two variables can be calculated [19]. Hence when comparing with simple linear regression, MLR perform better with less error rate. Moreover, multiple regression can be implemented in linear and non-linear modelling. Multiple regression is based on the statement that there is a linear relationship between both dependent and independent variables, where no assumption was made for major correlation between the independent variables [18-20].

B. Support Vector Regression

Support-vector regression (SVR) comes from a machine learning model namely the support-vector machine (SVM). SVM can work for regression tasks and is suggested in order to forecast tourism demand. SVM is a class of linear algorithms that can be used for classification, regression, density estimation, novelty detection, and other applications. SVM uses classification techniques to build a predictive model where its algorithm's main purpose is to find a hyperplane in an N-dimensional space that distinctly classify the data points. Separating two classes of data points may lead to many possible hyperplanes. The hyperplane equation is reflected in (2) and in Fig. 1:

$$\vec{w} \cdot \vec{x} + b = 0 \quad (2)$$

where w is a weight vector, x is input vector and b is bias. SVM searches for the hyperplane with the largest margin in separating the circle objects and square objects with minimum classification error.

An optimal network structure can be achieved by SVR on the basis of the theory of structural risk minimization which the margin of the hyperplane will be maximized [15, 21]. The main differences between Least Square Support Vector Machine (LSSVM) and SVM is that LSSVM includes the equality of constraints instead of the inequalities, and it is based on the least squares cost function [22]. SVR has been successfully implemented to solve forecasting problems in many fields, such as financial time series (stock index and exchange rate) forecasting, engineering and software (production values and reliability) forecasting, atmospheric science forecasting, electric load forecasting and commodity demand forecasting [16].

The SVR model has been successfully applied to forecast tourist arrivals too. Empirical research has shown that the choice of the parameters in an SVR model significantly influences the accuracy of forecasting. SVR solves the problems of estimation, classification, and nonlinearity via its loss function [17]. Tourism data often exhibit nonlinear characteristics, with SVR widely used in the tourism demand forecast. Low speed, however, is the key drawback of SVR in the training process [23], due to various hyperparameters setting in the model. Some inappropriate SVR parameters allow for the occurrences of overfitting or underfitting problem.

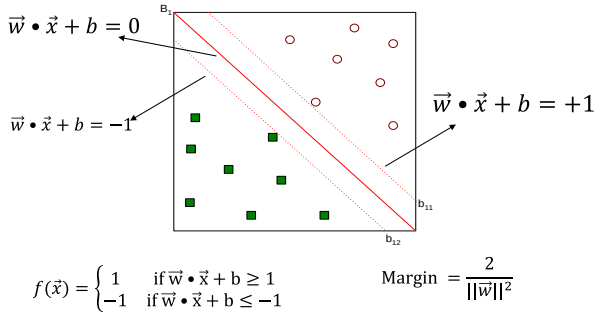

Fig. 1. The Hyperplanes for Two Classes of Objects [21].

C. Decision Tree Regression

Decision tree algorithm that has been used for classification or regression predictive modelling problems is called Classification and Regression Trees (CART). Decision Tree is one of the classifiers in supervised learning algorithm with a tree-like structure. It consists of root, interior, and leaf nodes in which the outcomes are represented at leaf node. CART is relatively straight-forward for prediction making. The algorithm works through multiple iterations until the tree is able to predict a proper value for the data point. Among the benefits of using CART algorithm is that it is easy to understand, less data cleaning process, non-linearity does not affect the output of the model, and the number of hyper-parameters to be tuned is almost null [21, 24]. The drawback is that it may have an overfitting problem, but which can be solved using the Random Forest algorithm.

The split attribute in the tree is chosen based on the standard deviation for the independent variable and dependent variable (outcome), and the formula for the standard deviation based on an attribute *x* as shown in equation (3):

$$S(x) = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n}} \quad (3)$$

where, S is standard deviation, x_i is the value of the attribute, $\overline{x}$ is the mean value of x and n is the record of x.

The standard deviation reduction is based on the decrease in standard deviation after a dataset is split on an attribute. Construction of a decision tree is basically finding an attribute that have the highest standard deviation reduction (SDR), where according to the calculation in equation (3) as the most homogeneous branch. In other words, the standard deviation of the target will be compared to different standard deviation of each independent variables in the dataset.

III. Data Source And Model Evaluation

A. Data Preparation

The data collection is obtained from different sources which are opened to both public and private use. This research study used monthly data points with observation period from year 2012 until 2018. The first dataset collected was from Google Trends, which is a search trends feature that displays the frequency with which a certain search term is entered into Google's search engine in relation to the site's total search volume over time. There are six keywords or search term used for this study: 1) "Taman Negara", 2) "taman negara accommodation", 3) "taman negara resort", 4) "taman negara canopy walkway", 5) "Cuti Cuti Malaysia", and 6) "visit Malaysia".

The second dataset was retrieved from the Federal Reserve Bank of St. Louis (FRED) where the dataset recorded monthly currency exchange rate with units of Malaysian Ringgit to One U.S. Dollar. The third dataset was extracted from Visual Crossing website that provides historical weather public data. To generate the worldwide weather observation database, the website processes millions of hourly weather observations from thousands of observation stations. However, the dataset extracted from the website is in weekly frequency. Hence, the need to calculate the monthly frequency of average attributes were done using the formula of average in Microsoft Excel. The attributes extracted from the source are Location, Date and Time, Maximum Temperature (degC), Minimum Temperature (degC), Temperature (degC), Heat Index (degC), Chance Precipitation (%), Precipitation (mm), Wind Speed (kph), Wind Direction, Wind Gust (kph), Visibility (km), Cloud Cover, Relative Humidity, and Conditions. However, for this research process, only two attributes were selected: 1) Temperature (degC); 2) Precipitation (mm).

The last and most important dataset is the total visitors' arrival to Taman Negara Pahang, Malaysia. This public dataset was found in a Malaysia Open Data Portal at data.gov.my. However, the data is in a yearly term basis and only covers data from the year 2012 until 2018. Thus, an additional data obtained from Jabatan Perlindungan Hidupan Liar dan Taman Negara (PERHILITAN) Pahang was used to identify the monthly number of visitors to Taman Negara Pahang. Since the data collected for this research were huge and obtained from different sources hence making it unstructured, data transformation process needed to be done. Data preprocessing is the process of changing the variety of raw dataset into one dataset suitable to be used in software such as RapidMiner. For this study, the regression modelling was done with the help of data analytics software tool, RapidMiner. Fig. 2 shows the conceptual framework for the regression modelling that illustrated the source of data, independent variables, and dependent variables. The variables for the observation period (year 2012-2018) were Date, Currency Exchange, Visit Malaysia, Cuti Cuti Malaysia, Taman Negara, Taman Negara accommodation, Taman Negara Resort, Taman Negara canopy walkway, Average Temperature, Average Precipitation and Total of visitors.

B. Model Evaluation

For this study, a comparison was made between three predictive models, namely Multiple Linear Regression (MLR), Support-Vector Regression (SVR), and Decision Tree Regression (DTR), as shown in Fig. 3. Among the important tasks conducted were data preparation from four data sources: data preprocessing, training and testing data split, modeling with three algorithms, and finally evaluation and deployment. The model development in this research study used RapidMiner's default values. The comparison was measured based on the models' performances by manipulating the sampling type and number of folds in cross validation.

The RMSE (Root Mean Squared Error) is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. While squared_correlation is a relative measure of fit, RMSE is an absolute measure of fit. Thus, the lower the values of RMSE, the better it is. The formula for Root Mean Square Error (RSME) is as in (4),

$$\sqrt{\sum_{i=b}^{n}(X_{actual} - X_{model})^2} \quad (4)$$

where, x_{actual} is an observed value and x_{model} is the predicted value.

Source	Independent Variables		Dependent Variables
FRED	Currency Exchange		
Google Trends	Keywords: Visit Malaysia, Cuti Cuti Malaysia, Taman Negara, Taman Negara accommodation, Taman Negara Resort, Taman Negara canopy walkway,	Regression Model	Total of visitors
Visual Crossing	Weather: Average Temperature, Average Precipitation		

[a] FRED: https://fred.stlouisfed.org/categories/15

[b] Visual Crossing: https://www.visualcrossing.com/

Fig. 2. Conceptual Framework of Tourist Visit Regression Model.

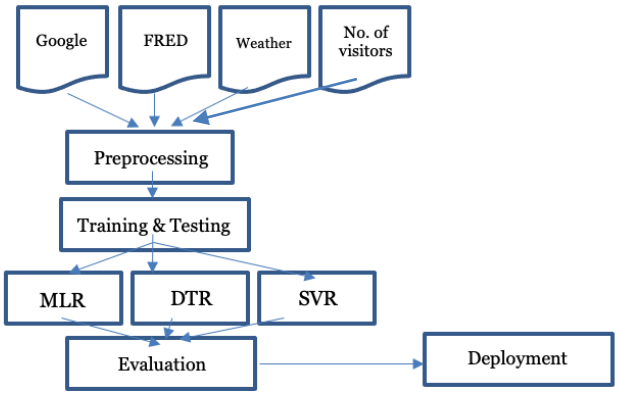

Fig. 3. The Overview of Processes Involved in the Study.

Another useful measure to determine the optimal model is goodness-of-fit measurements, such as the square of the correlation coefficient (r^2 or squared_correlation). This measure is used to examine the scattered-locations of data points around the fitted value. The number denotes the percentage of variance in one variable that can be explained by the other. The higher the r^2 score, the more precise the prediction. It does not, however, tell how precise the forecasts are in the dependent variable's units.

IV. Result and Discussion

This section presents the results and provides a discussion based on the outlined framework given in Fig. 2 and Fig. 3.

A. Data Description

The final dataset was constructed based on the following data sources:

- The Google Trends dataset, in Fig. 4 recorded 504 data which were generated from monthly frequency of year 2012 until year 2018 for six related keywords, as mentioned in section 3 (12 months/year × 7 years × 6 keywords = 504 data).

	A	B	C	D	E
1		Category: All categories	Category: All categories	Category: All categories	Category: All categories
2	Month	Taman Negara: (Malaysia)	taman negara accommodation: (M	taman negara resort: (Malaysia)	taman negara canopy walkway: (Mala
3	2012-01	39	0	39	0
4	2012-02	64	0	20	0
5	2012-03	64	0	26	0
6	2012-04	45	0	35	0
7	2012-05	59	0	75	0
8	2012-06	57	0	42	0
9	2012-07	50	0	16	0
10	2012-08	44	69	60	0
11	2012-09	49	0	16	56
12	2012-10	45	59	15	0
13	2012-11	38	0	7	0
14	2012-12	37	0	35	0
15	2013-01	48	57	35	0
16	2013-02	44	0	16	0
17	2013-03	32	0	32	0
18	2013-04	48	0	19	0
19	2013-05	43	0	18	0
20	2013-06	37	0	38	0
21	2013-07	41	0	30	0
22	2013-08	39	51	32	0

Fig. 4. Google Trend Hits.

- Second dataset, retrieved from FRED (Federal Reserve bank of St. Louis) - the dataset recorded monthly currency exchange rate with units of Malaysian Ringgit to One U.S. Dollar, for each month from 2012, as in Fig. 5.

10	Frequency: Monthly	
11	observation_date	EXMAUS
12	2012-01-01	3.1092
13	2012-02-01	3.0220
14	2012-03-01	3.0444
15	2012-04-01	3.0586
16	2012-05-01	3.0978
17	2012-06-01	3.1783
18	2012-07-01	3.1653
19	2012-08-01	3.1153
20	2012-09-01	3.0758
21	2012-10-01	3.0524
22	2012-11-01	3.0555
23	2012-12-01	3.0537
24	2013-01-01	3.0407
25	2013-02-01	3.0964
26	2013-03-01	3.1074
27	2013-04-01	3.0480
28	2013-05-01	3.0188
29	2013-06-01	3.1433

Fig. 5. Daily Currency Exchange.

- Third dataset, extracted from Visual Crossing website which provides historical weather public data. To generate the worldwide weather observation database, the website processes millions of hourly weather observations from thousands of observation stations. However, the dataset extracted from the website is in weekly frequency.

- Fourth dataset consists of the number of visitors in Taman Negara National Park, including nearby places and Gunung Tahan from Malaysia Open Data Portal by yearly and PERHILITAN Pahang, by monthly, starting from January till December, as in Table I. The highest number of visitors found was in March.

Based on the gathered information, the final dataset consists of 11 variables or attributes with 84 rows of monthly data points for the observation period (year 2012-2018). The variables/columns are observation date (month), Currency exchange, the six keywords: Visit Malaysia, Cuti Cuti Malaysia, Taman Negara, Taman Negara accommodation, Taman Negara Resort and Taman Negara canopy walkway, Average Temperature, Average Precipitation and Total Visitors. The Fig. 6 shows the snippets of 10 rows from 2012 as the sample of the real dataset used in the study.

TABLE I. THE VISITORS OF TAMAN NEGARA

No	Month	Nearby places		Gunung Tahan		Total
1	January	991	3	58	-	**1052**
2	February	1347	6	50	-	**1403**
3	Mac	1850	46	55	18	**1969**
4	April	823	28	119	-	**970**
5	May	1260	7	106	12	**1385**
6	June	1296	23	157	35	**1511**
7	July	855	19	12	2	**888**
8	August	1479	40	112	-	**1631**
9	September	1387	20	125	-	**1532**
10	October	1361	13	188	-	**1562**
11	November	1080	-	5	-	**1085**
12	December	574	-	-	-	**574**
		Total				**15562**

Open Data: https://www.data.gov.my/

observation_da	Currency exhange	visit malaysia: (	Cuti Cuti Malaysia (	Taman Negara:	taman negara acco	taman negar	taman negara c	Average Tempe	Average Preci	Total Visitors
1/1/2012	3.1092	31	39	39	0	39	0	25.44194	15.59032	6722
1/2/2012	3.022	31	36	64	0	20	0	26.44483	2.327586	8642
1/3/2012	3.0444	29	26	64	0	26	0	26.47097	7.383871	11523
1/4/2012	3.0586	36	30	45	0	35	0	26.92	12.03	5761
1/5/2012	3.0978	32	37	59	0	75	0	27.31613	11.03226	8642
1/6/2012	3.1783	31	33	57	0	42	0	27.68667	5.383333	9602
1/7/2012	3.1653	38	30	50	0	16	0	27.12903	3.383871	5761
1/8/2012	3.1153	35	27	44	69	60	0	26.9871	9.070968	9602
1/9/2012	3.0758	44	37	49	0	16	56	27.03333	5.286667	9602
1/10/2012	3.0524	33	41	45	59	15	0	26.47097	5.829032	9602

Fig. 6. Sample of Merged Dataset.

Meanwhile Fig. 7 to 10 show statistical measures of selected attribute. In Fig. 6, the currency exchange is represented in real value, while for the six keywords (Visit Malaysia, Cuti Cuti Malaysia, Taman Negara, Taman Negara accommodation, Taman Negara Resort and Taman Negara canopy walkway) are represented as the count of the words mentioned every month. The count of word "Taman Negara accommodation" being mentioned were found to be more as compared to other words. The total number of word count in the dataset is as shown in Fig. 10, with "Visit Malaysia" being the highest count and "Taman Negara canopy walkway" as the lowest count.

Name	Type	Missing	Statistics			Filter (11 / 11 attributes):
Currency exhange	Real	0	Min 3.019	Max 4.457	Average 3.697	
visit malaysia: (Worldwide)	Integer	0	Min 25	Max 85	Average 44.202	
Cuti Cuti Malaysia (Worldwide)	Integer	0	Min 16	Max 57	Average 30.048	
Taman Negara: (Malaysia)	Integer	0	Min 29	Max 67	Average 44.083	
taman negara accommodation: ...	Integer	0	Min 0	Max 100	Average 11.464	
taman negara resort: (Malaysia)	Integer	0	Min 5	Max 75	Average 24.024	
taman negara canopy walkway: ...	Integer	0	Min 0	Max 75	Average 6.786	

Fig. 7. Statistical Measures for Currency Exchange Rate and Count of the Keywords Searched in Google.

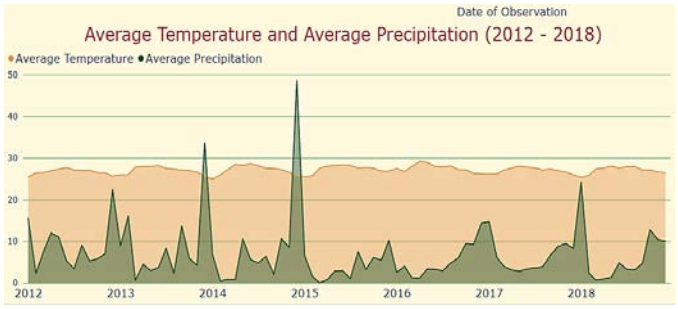

Fig. 8. The Graph for Average Temperature and Average Precipitation from 2012 Till 2018.

In Fig. 8, the average temperature and average precipitation are presented from 2012 till 2018, with gathered information from Visual Crossing. The average temperature is between 25 and 30 degrees Celsius. Here, the precipitation shows the highest value in 2015 and the lowest are shown to be present in every year, due to the dry season that occurs in the respective years.

Fig. 9 illustrates the graph of total visitors in every month to the Taman Negara from 2012 till 2018. The total number of visitors continuously changed over time.

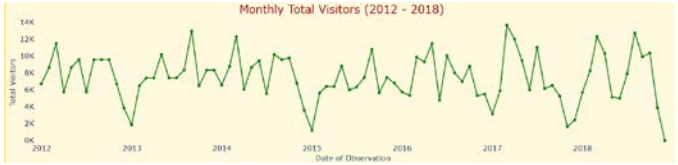

Fig. 9. The Graph for Monthly Total Visitors from 2012 Till 2018.

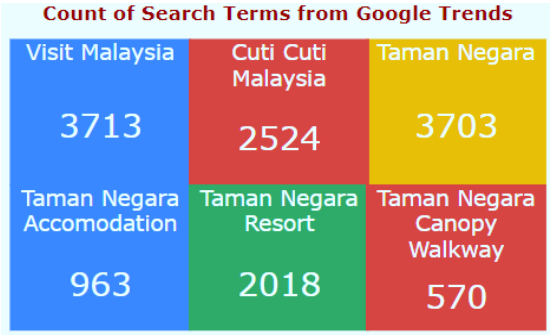

Fig. 10. Count of Words related to Taman Negara.

A total of 30 experiments were performed by tuning five different numbers of folds with two types of sampling in three predictive models (5×2×3 = 30). The comparison table shows the overall performance recorded during the experiment for the modelling.

B. Multiple Linear Regression Results

The best model, multiple linear regression, resulted with the RMSE values of 2545.977 and the values of r^2 is 0.276 which is the highest value with linear sampling among four models, as shown in Table II. Thus, linear regression equation for the best model is as the following:

*Predicted (Total Visitors) = - 631.866 * Currency exchange*
*+ 7.229 * visit malaysia: (Worldwide)*
*- 9.240 * Cuti Cuti Malaysia (Worldwide)*
*+ 113.991 * Taman Negara: (Malaysia)*
*- 7.465 * taman negara accommodation: (Malaysia)*
*- 2.319 * taman negara resort: (Malaysia)*
*+ 5.474 * taman negara canopy walkway: (Malaysia)*
*+ 1062.211 * Average Temperature*
*- 20.455 * Average Precipitation*

TABLE II. LINEAR REGRESSION MODELLING

Sampling method	**RMSE**	r^2
Linear Sampling (k = 10)	2624.972	0.301
Linear Sampling (k = 8)	2545.977	0.276
Shuffled Sampling (k = 10)	2570.164	0.227
Shuffled Sampling (k = 4)	2591.315	0.162

C. Decision Tree Regression Results and Rules

Modeling of decision tree resulted in a set of rules represented in a tree-like structure. Each node corresponds to a splitting rule for a single attribute. Fig. 11 shows the extracted tree and the conditions of each predicted total of tourist visit. Table III shows the result of RMSE and r^2 for the decision tree regression, in which the RMSE is higher as compared to multiple linear regression. As can be seen, average temperature and the count of the keywords appeared to be among the important variables.

```
Average Temperature > 26.984
|   Currency exhange > 4.371: 12859.000 (count=2)
|   Currency exhange ≤ 4.371
|   |   Taman Negara: (Malaysia) > 49.500
|   |   |   taman negara accommodation: (Malaysia) > 38: 6295.000 (count=2)
|   |   |   taman negara accommodation: (Malaysia) ≤ 38
|   |   |   |   visit malaysia: (Worldwide) > 40
|   |   |   |   |   Average Temperature > 27.417
|   |   |   |   |   |   Currency exhange > 4.070: 11782.500 (count=2)
|   |   |   |   |   |   Currency exhange ≤ 4.070
|   |   |   |   |   |   |   Currency exhange > 3.408: 8386.500 (count=2)
|   |   |   |   |   |   |   Currency exhange ≤ 3.408: 9907.000 (count=2)
|   |   |   |   |   Average Temperature ≤ 27.417
|   |   |   |   |   |   taman negara canopy walkway: (Malaysia) > 16: 12665.500 (count=2)
|   |   |   |   |   |   taman negara canopy walkway: (Malaysia) ≤ 16: 11138.000 (count=2)
|   |   |   |   visit malaysia: (Worldwide) ≤ 40
|   |   |   |   |   visit malaysia: (Worldwide) > 35: 6377.000 (count=2)
|   |   |   |   |   visit malaysia: (Worldwide) ≤ 35: 9122.000 (count=2)
|   |   Taman Negara: (Malaysia) ≤ 49.500
|   |   |   visit malaysia: (Worldwide) > 49.500
|   |   |   |   taman negara resort: (Malaysia) > 32.500: 9070.500 (count=2)
|   |   |   |   taman negara resort: (Malaysia) ≤ 32.500
|   |   |   |   |   Taman Negara: (Malaysia) > 44.500: 6309.500 (count=2)
|   |   |   |   |   Taman Negara: (Malaysia) ≤ 44.500
```

Fig. 11. Rules Extracted from the Decision Tree Regression.

TABLE III. DECISION TREE REGRESSION MODELLING RESULTS

Sampling Method	**RMSE**	r^2
Linear Sampling (k = 4)	3278.578	0.07
Linear Sampling (k = 9)	3474.772	0.104
Shuffled Sampling (k = 4)	3424.049	0.096
Shuffled Sampling (k = 9)	3570.849	0.058

Based on the lowest RSME, the best SVR is when linear sampling done with k = 4, at 3278.578, though the r^2 is the lowest among the other experiments.

D. Support Vector Regression Results

The Support Vector Regression produces an average result between Multiple Linear Regression and Decision Tree Regression. Table IV displays the best result for linear sampling and shuffled sampling. Based on the lowest RSME, the best SVR is at linear sampling with k = 9, at 2727.532 and the r^2 is the second best among the experiments conducted.

TABLE IV. SUPPORT VECTOR REGRESSION MODELLING RESULTS

Sampling Method	**RMSE**	r^2
Linear Sampling (k = 9)	2727.532	0.286
Linear Sampling (k = 10)	2749.607	0.306
Shuffled Sampling (k = 9)	2741.599	0.234
Shuffled Sampling (k = 10)	2741.599	0.234

E. Best Model Deployment

Selection of best model is measured by the lowest RMSE value and the highest value of the squared correlation between the predicted and the actual values. The experiment proved that the decision tree model displayed a weak performance for this research study as it produced a greater error as compared to other model at any parameters tuning. Between MLR and SVR, the error produced in MLR is almost the lowest, but the squared correlation value is lesser than the SVR model. Additionally, the cross-validation with number of folds, k = 8 for the linear sampling type indicated the best model is MLR which outperforms SVR and Decision Tree. The assumption is made that the number of folds is depending on the number of instances in the dataset and the sampling type is based on the problem model, which in this study the input is the linear problem to find the relationship between these variables.

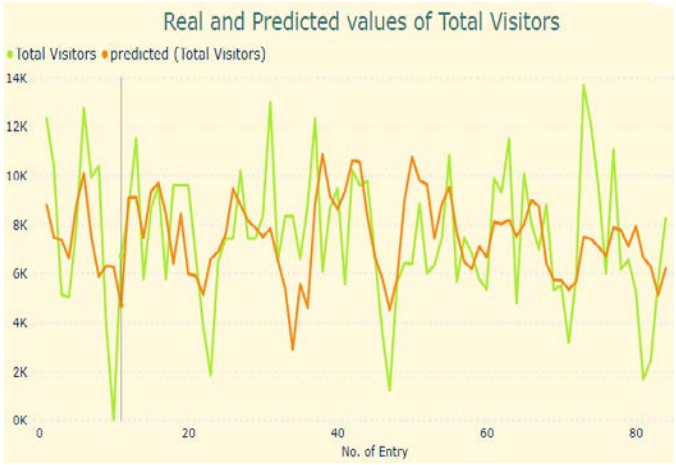

Fig. 12. The Graph of Real (Actual) Value and Predicted Value of Total Visitors.

Fig. 13. The Graph of Prediction until 2030.

Fig. 12 represents the real (actual) value of visitors and the predicted value of visitors by using multiple linear regression. The advantage of using multiple linear regressions is that when given values in decimal point, the results can be easily interpreted by the decision maker. The graph also visualizes clearly the gap between actual and predicted value of the visitors. Subsequently, Fig. 13 shows the predicted total number of visitors until 2030 with some simulated values for each variable in the best MLR model. In future, this research can be extended by providing the estimated values for each relevant variable, and the predicted total number of visitors can then be stipulated to the tourist management.

V. Conclusion

This paper presented the implementation of regression models, namely Multiple Linear Regression (MLR), Support Vector Regression (SVR) and Decision Tree Regression (DTR). A set of variables were constructed based on the selected keywords, the currency exchange and the weather variables for predicting the number of total visitors. The experiments conducted had indicated that these regression algorithms were able to predict the total number of visitors to Taman Negara National Park. The results for the experiments after tuning of parameters demonstrated an improved accuracy of the models since it can control the complexity, which indirectly prevented from overfitting of the model. In this study, the linear problem (input) discovered was to find the relationship between the factors affecting the demand for the total number of visitors to Taman Negara National Park.

Multiple Linear Regression model with linear sampling type and 8-fold cross validation approach appeared to be the best model. The experiments showed that the best parameters setting was based on the instances of the dataset itself. Consequently, some suggestions for future works to improve the quality of the research study were identified. Firstly, the use of advanced visualization tools to work with real-time data to the dashboard can be applied. Next, more data ought to be collected to produce a better performance of predictive models, such as different keywords and any other related campaigns. Lastly, the use of hybrid machine learning and optimization algorithms can be considered to optimize the parameter tuning for better accuracy. The developed model is useful to the tourism management, for predicting the number of visitors to Taman Negara National Park, Malaysia. The tourism management as the user can improve their operations by making a strategic decision making based on the predicted outcomes. If the tourism destination can operate more smoothly, the visitors can reap the benefits from the meaningful experience they received when visiting the national park. Other study can also be performed such as the length of stay and recommended activities based on the tourists' profiles.

Acknowledgment

The authors would like to thank the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia for all the support given.

References

[1] G. Ghani, N. Mohamad, and M. I. Ariffin, "Malaysia's Tourism Demand: A Gravity Model Approach (Permintaan Pelancongan Di Malaysia: Pendekatan Model Graviti)," vol. 6, pp. 39-50, 03/01 2018.

[2] A. Shuib, "Tourism in Taman Negara Malaysia Its Contribution as Perceived by Residents of Ulu Tembeling," Akademika, vol. 47, 1995.

[3] "National Park (Taman Negara) of Peninsular Malaysia." https://whc.unesco.org/en/tentativelists/5927/ (accessed 13 July 2021, 2021).

[4] Z. Samdin, Y. A. Aziz, and M. R. Yacob, "Ecotourism in Taman Negara National Park: issues and challenges," 2013.

[5] H. Reduan, "New Straits Times: Protecting Taman Negara," 22 June 2017 2017. [Online]. Available: https://www.nst.com.my/opinion/columnists/2017/06/251183/protecting-taman-negara.

[6] A. Aziz, S. Nur, M. Jamaludin, N. Idris, and M. Mariapan, "The Attractiveness Of Taman Negara National Park, Malaysia As Perceived By Local Visitors," vol. 33, pp. 1-13, 12/15 2018.

[7] "Corporate Finance Institute: Regression Analysis." https://corporatefinanceinstitute.com/resources/knowledge/finance/regression-analysis/ (accessed 26 November 2021.

[8] M. Hanafiah and M. Harun, "Tourism Demand in Malaysia: A cross-sectional pool time-series analysis," International Journal of Trade, Economics and Finance, vol. 1, pp. 80-83, 01/01 2010, doi: 10.7763/IJTEF.2010.V1.15.

[9] S. S. A. Kosnan, N. Ismail, and S. Kaliappan, "Determinants of international tourism in malaysia: Evidence from gravity model," Jurnal Ekonomi Malaysia, vol. 47, pp. 131-138, 01/01 2013.

[10] W. Höpken, T. Eberle, M. Fuchs, and M. Lexhagen, "Google Trends data for analysing tourists' online search behaviour and improving demand forecasting: the case of Åre, Sweden," Information Technology and Tourism, Article vol. 21, no. 1, pp. 45-62, 2019, doi: 10.1007/s40558-018-0129-4.

[11] K. Li, W. Lu, C. Liang, and B. Wang, "Intelligence in tourism management: A hybrid FOA-BP method on daily tourism demand forecasting with web search data," Mathematics, Article vol. 7, no. 6, 2019, Art no. 531, doi: 10.3390/MATH7060531.

[12] U. Claude, "Predicting tourism demands by google trends: A hidden markov models based study," Journal of System and Management Sciences, Article vol. 10, no. 1, pp. 106-120, 2020, doi: 10.33168/JSMS.2020.0108.

[13] A. H. B. Pengiran Bagul, Developing Climate Change Dimensions In Malaysia Through Tourists' Perception. 2013.

[14] S. Cankurt and A. Subasi, "Tourism demand modelling and forecasting using data mining techniques in multivariate time series: A case study in Turkey," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 24, 01/01 2015, doi: 10.3906/elk-1311-134.

[15] C. Zhong-jian, L. Sheng, and Z. Xiao-bin, "Tourism demand forecasting by support vector regression and genetic algorithm," in 2009 2nd IEEE International Conference on Computer Science and Information Technology, 8-11 Aug. 2009 2009, pp. 144-146, doi: 10.1109/ICCSIT.2009.5234447.

[16] W.-C. Hong, Y. Dong, L.-Y. Chen, and S.-Y. Wei, "SVR with hybrid chaotic genetic algorithms for tourism demand forecasting," Applied Soft Computing, vol. 11, no. 2, pp. 1881-1890, 2011/03/01/ 2011, doi: https://doi.org/10.1016/j.asoc.2010.06.003.

[17] Z.-y. Mei, H. Qiu, C. Feng, and Y. Cheng, "Research on a forecasting model of tourism traffic volume in theme parks in China,"

Transportation Safety and Environment, vol. 1, no. 2, pp. 135-144, 2019, doi: 10.1093/tse/tdz011.

[18] A. Hayes. "Multiple Linear Regression (MLR) definition." Investopedia. https://www.investopedia.com/terms/m/mlr.asp (accessed 26 November 2021.

[19] E. Sreehari and S. Srivastava, "Prediction of Climate Variable using Multiple Linear Regression," in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 14-15 Dec. 2018 2018, pp. 1-4, doi: 10.1109/CCAA.2018.8777452.

[20] M.S. Hilmi, S. Mutalib, S.R. Sharif and S.N.K. Kamarudin, "Forecasting electricity demand from daily log sheet with correlated variables," ESTEEM Academic Journal 16, 31-41.

[21] P-N. Tan, M. Steinbach, A. Karpatne, V. Kumar, "Introduction to Data Mining, 2nd Edition," 2019.

[22] M.Z. Shahrel, S. Mutalib, S. Abdul-Rahman, "PriceCop-Price Monitor and Prediction Using Linear Regression and LSVM-ABC Methods for E-commerce Platform," International Journal of Information Engineering & Electronic Business, vol. 13 (1), 2021, pp. 1-14, doi: 10.5815/ijieeb.2021.01.01.

[23] F.-C. Yuan, "Intelligent forecasting of inbound tourist arrivals by social networking analysis," Physica A: Statistical Mechanics and its Applications, vol. 558, p. 124944, 2020/11/15/ 2020, doi: https://doi.org/10.1016/j.physa.2020.124944.

[24] N.A.M. Salim, Y.B. Wah, C. Reeves et al. Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques. Sci Rep 11, 939 2021, https://doi.org/10.1038/s41598-020-79193-2.

[25] R. Sovia, M. Yanto, Yuhandri, "Prediction Tourist Visits With Multiple Linear Regressions in Artificial Neural Networks," Turkish Journal of Computer and Mathematics Education, vol. 12 No. 3, pp. 1492-1501, 2021.

[26] N. Agustina, C.D. Puspita, D. Arifatin, R. Yordani, "Application of Logistic Regression to Determine The Quality of Foreign Tourists to Indonesia," In Journal of Physics: Conference Series 2021 Mar 1 (Vol. 1863, No. 1, p. 012029). IOP Publishing.

[27] M. Jackman and S. Naitram, "Segmenting tourists by length of stay using regression tree models", Journal of Hospitality and Tourism Insights, Vol. ahead-of-print No. ahead-of-print, 2021 https://doi.org/10.1108/JHTI-03-2021-0084.

[28] Nurkhodzha Akbulaev, Gulnar Mirzayeva, "Analysis of a paired regression model of the impact of income from international tourism on the foreign trade balance," African Journal of Hospitality, Tourism and Leisure, Volume 9(1), pp 1-13, 2020.

[29] R. Zhang, "Exploration of Social Benefits for Tourism Performing Arts Industrialization in Culture-tourism Integration based on Deep Learning and Artificial Intelligence Technology," Frontiers in Psychology. 2021;12:37.

[30] S. Moghtasedi, C. I. Muntean, F. M. Nardini, R. Grossi and A. Marino, "High-Quality Prediction of Tourist Movements using Temporal Trajectories in Graphs," In 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 348-352, 2020.

[31] F. Qian, C. Han and M. Haiyan, "Intelligent model system for tourism flow prediction: a study of Xi'an Museum," In Proceedings of the 2016 International Conference on Intelligent Information Processing, pp. 1-7, 2016.