Jirapond Muangprathub / Patthamaphon Kaewmanee / Jarunee Saelee / Pattaraporn Warintarawej / Wichuta Sae-jie

Forecasting Trends in Foreign Tourism by Machine Learning

Abstract

Tourists frequently use online search engines for travel planning, making search data a valuable predictor of future tourism volume. This study employs machine learning to analyse the predictive power of keyword search data for forecasting tourist arrivals, incorporating a lag time between searches and arrivals. The dataset is collected and prepared from two sources: a search engine and government agencies, covering the years 2014-2019, to be analysed by machine learning. The SARIMA model effectively forecasts trends in keyword searches and tourist numbers, while SVM (Support Vector Machine) and Random Forest outperform other methods in predicting arrivals. This research supports tourism operators and stakeholders in planning for future tourists, utilising the obtained keywords to enhance visibility in tourist searches through SEO.

Keywords: tourism volume forecasting, tourist arrivals prediction, machine learning, search engine data, tourist data analysis

1. Introduction

In the digital era, online search engines have become essential tools for travel planning, with tourists frequently relying on them to explore destinations, accommodations, and attractions. The widespread use of search engines has made search data a valuable source for predicting future tourism demand (Zervas et al., 2017; Cheng et al., 2018). The present study focuses on using keyword search data from Google Trends to forecast tourist arrivals. By analysing the relationship between search volumes and actual tourist numbers, this research seeks to improve forecasting accuracy and provide valuable insights for stakeholders in the tourism industry. In Thailand, the number of visitors has shown remarkable growth, increasing from 14.1 million in 2009 to 38.3 million in 2018 (Adulwattana et al., 2019). This surge highlights the growing significance of tourism as a key driver of Thailand's GDP growth. In 2019, the tourism sector played a pivotal role in the Thai economy, generating 86,908 million dollars (3.01 trillion baht) in revenue from 39.7 million foreign tourists. The country's exceptional tourism potential has contributed to its attractiveness as a destination for international visitors (Wongsathan et al., 2018). The tourism sector is financially strong and has achieved good results, with effects that extend across various sectors, including restaurants, large and small shops, farmers, and transportation businesses. Planning marketing strategies for entrepreneurs in the tourism industry is essential to attract tourists (Kaewmanee et al., 2021). The search engine tool is primarily used to support travel

Jirapond Muangprathub, PhD, Associate Professor, Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Surat Thani, Thailand; ORCID ID: https://orcid.org/0000-0002-3062-4696; e-mail: jirapond.m@psu.ac.th

Patthamaphon Kaewmanee, MSc, Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Surat Thani, Thailand; e-mail: 6240320502@psu.ac.th

Jarunee Saelee, PhD, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani, Thailand; ORCID ID: https://orcid.org/0000-0002-4925-4170; e-mail: jarunee.sa@psu.ac.th

Pattaraporn Warintarawej, PhD, Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Surat Thani, Thailand; ORCID ID: https://orcid.org/0000-0002-2034-1932; e-mail: pattaraporn.w@psu.ac.th

Wichuta Sae-jie, PhD, Corresponding Author, Assistant Professor, Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Surat Thani, Thailand; ORCID ID: https://orcid.org/0000-0001-7056-7529; e-mail: wichuta.sa@psu.ac.th

Original Research Article
Vol. 73/ No. 3/ 2025/ 392 - 409
UDC: 338.486.5:81'322; https://doi.org/10.37741/t.73.3.1

plans by providing information before detailed planning (Asaithambi et al., 2023). Thus, search engines can inform tourism operators about tourist interests, helping them anticipate changes in demand and plan for meeting future expectations.

With advances in online access to the web, search engines have become a necessity for potential travellers who want to plan their trips, enabling them to search for accommodations, climates, tourist attractions, and more. Thus, the keyword searches are associated with accurate future demand data. In prior studies, data from search engines have been utilised in numerous forecasts for the tourism industry, demonstrating their ability to predict tourist volumes (Bi et al., 2020; Feng et al., 2019; Liu et al., 2019; Xie et al., 2021). Machine learning is one approach that can be applied to tourism data analysis (Law et al., 2019; Xie et al., 2021; Sun et al., 2019). In the previous studies, the focus has been on applying only one machine learning method. However, each technique in machine learning provides a different match with the nature of the data. Thus, this work compares alternative machine learning methods to choose the best alternative for predictions.

The data were derived from Google Trends and the Ministry of Tourism and Sports, Thailand, for the years 2014-2019, respectively, for keyword search index data and for the volume of foreign tourists. The study area focused on the islands of Koh Samui, Koh Phangan, and Koh Tao in Surat Thani Province, Thailand. Koh Samui was ranked seventh in the world and second in Asia among the best islands for tourism in 2021 by Travel + Leisure magazine, a renowned global travel magazine in the United States (Sachdev 2021; Asher-Walsh 2021). These islands were used to study the relationship between keyword searches by potential tourists and their subsequent arrival rates. The statistical and machine learning methods employed included time series analysis, Artificial Neural Network (ANN), Deep Learning, Random Forest, and Support Vector Machine (SVM) to identify the most predictive model for tourist volume and to analyse the search keywords associated with actual tourist demand. As its benefits, this study can support stakeholder planning, such as strategic planning or marketing by service providers to tourists, to mitigate future impacts of rapid changes in tourism (currently impeded by a pandemic) and to meet the increasing demands of tourism. The correlated or predictive keywords can also be utilised in online marketing, specifically in Search Engine Optimisation (SEO), by entrepreneurs using online media to enhance their potential to attract website visitors and sell more products and services.

2. Related prior studies

The searches conducted by potential tourists have shifted primarily from large display screens on personal computers to mobile devices, which are always readily available. Search engines are valuable tools for travel-related data searches, particularly during the pre-trip planning stages. Moreover, searches of hotel data were affected by a lag period (Wickramasinghe et al., 2020). Internet searches have been widely used in tourism-related research, and processes for forecasting with Internet data have been introduced (Li et al., 2021). The types of models used in predicting include time series, artificial intelligence, and deep learning models. Among time series models based on historical data, the SARIMA model is one of the most widely used for forecasting tourism demand. Forecasting the monthly number of foreigners visiting India, the Naive I and Naive II models, as well as the seasonal autoregressive integrated moving average (SARIMA) and Grey models, were demonstrated using performance measures such as the mean absolute percentage error (MAPE), U-statistic, and turning point analysis (TPA) (Chandra et al., 2018). Another study examined the forecasting performance of two nonlinear computational methods—artificial neural networks and genetic programming—by analysing monthly international tourism demand in Spain. The results indicate that these nonlinear methods provide slightly better predictions compared to the traditional SARIMA model. However, the improvement was only nearly statistically significant, raising questions about whether the higher implementation costs are justified (Alvarez et al. 2019). Forecasting hotel room demand was demonstrated using data from Google Trends and time series analysis (Pan et al., 2012). Predictions are crucial for the tourism industry in making

informed decisions and preparing significant investments. The relationship between social media and environmental data has been explored through time series analysis (Khatibi et al., 2020). Tourism management predicts daily tourism demand at the Huangshan scenic spot in China by filtering keywords from the Baidu index and utilising a backpropagation (BP) neural network model optimised with a fruit fly optimisation algorithm (FOA) (Li et al., 2019).

Next, the training algorithm used with an artificial neural network (ANN) is a significant determinant of its effectiveness and efficiency. Several algorithms have been compared on various benchmark problems (Yaghini et al. 2013). Forecasting monthly tourist volume in Hong Kong from mainland China utilised data on Baidu index terms, employing a technique that combines linear and nonlinear component models (Wen et al., 2019) and a Mixed Data Sampling Model (Wen et al., 2021). Wu et al. (2021) set as their forecasting target the daily tourist arrivals in Macau, China, using data from Google Trends and the Baidu index in a seasonal autoregressive integrated moving average (SARIMA) model combined with a long short-term memory (LSTM) model, known as SARIMA-LSTM. Additionally, Law et al. (2019) investigated tourist volume in Macau, China, using data from Google Trends and the Baidu index, employing a deep learning approach. Peng et al. (2021) predicted the volume of tourism in Jiuzhaigou and Beijing, China, using Baidu index data in combination with random forest and long short-term memory models.

Previous studies on predicting tourism demand have focused on using internet data. According to reviews of tourism demand forecasting, there are four key topics related to the subject. The first topic is tourism since the objective of this work is to predict trends in foreign tourist volume. The second one is tourism demand forecasting with various techniques. The third one is the source of data, which is a search engine. The last topic is also essential for prediction, namely the choice of a machine learning model. A summary of prior studies is shown in Table 1.

Table 1
Summary of previous studies relevant to the current research

Previous research			The topic focus in the field of this research			
Title	**Year**	**Authors**	**Tourism**	**Forecasting tourist demand**	**Search engine**	**Machine learning model**
Forecasting hotel room demand using search engine data	2012	Pan, B., Wu, D. C., & Song, H.	✓	✓	✓	✓
Forecasting foreign tourist arrivals in India using time series models	2018	Chandra, S., & Kumari, K.	✓			✓
SARIMA intervention based forecast model for visitor arrivals to Chiang Mai, Thailand	2018	Wongsathan, R.	✓			✓
Forecasting international tourism demand using a non-linear autoregressive neural network and genetic programming	2019	Álvarez-Díaz, M., González-Gómez, M., & Otero-Giráldez, M. S.	✓	✓		✓
Tourism demand forecasting: A deep learning approach	2019	Law, R., Li, G., Fong, D. K. C., & Han, X.	✓	✓	✓	✓
Intelligence in tourism management: a hybrid FOA-BP method on daily tourism demand forecasting with web search data	2019	Li, K., Lu, W., Liang, C., & Wang, B.	✓	✓	✓	✓
Forecasting tourism demand using search query data: A hybrid modelling approach	2019	Wen, L., Liu, C., & Song, H.	✓	✓	✓	✓
Fine-grained tourism prediction: Impact of social and environmental features	2020	Khatibi, A., Belém, F., da Silva, A. P. C., Almeida, J. M., & Gonçalves, M. A.	✓	✓	✓	✓
Forecasting tourism demand with multisource big data	2020	Li, H., Hu, M., & Li, G.	✓	✓		✓
Review of tourism forecasting research with internet data	2021	Li, X., Law, R., Xie, G., & Wang, S.	✓	✓	✓	✓
The role of disaggregated search data in improving tourism forecasts: Evidence from Sri Lanka	2021	Wickramasinghe, K., & Ratnasiri, S.	✓		✓	✓

3. Data sources and data preparation

This section describes the data sources and preparation to validate and verify the proposed analytical model. The data were derived from two sources: a search engine and the Ministry of Tourism and Sports, Thailand. The datasets were prepared for subsequent analysis by machine learning.

3.1. Data sources

The number of tourist arrivals in Thailand has increased rapidly, from 14.1 million in 2009 to 38.3 million in 2018, driving the country's gross domestic product (GDP) growth. Consequently, the tourism industry in Thailand has garnered a high level of attention, with a focus on promoting tourism further and satisfying tourists through planned marketing strategies for entrepreneurs. Potential tourists use online media for travel planning, often by searches using a search engine. Search engines are thus capable of informing tourism operators about tourists' interests and anticipated changes in demand, helping operators plan for and meet future expectations. Google is the most popular search engine in the world (Dinis et al. 2019), especially among tourists visiting Thailand (https://www.similarweb.com/engines/thailand (2023)). It also provides a free service for historical search engine query volume data. Google has a search analysis tool, Google Trends, which can extract those data for analysis. Google Trends (https://trends.google.co.th (2023)) provides Google search data from January 2014 to December 2019. It reports a query index, which displays the frequency with which a search query has been searched relative to the total search-volume across different areas and languages. This reflects the popularity of a particular query and, thereby, the users' interests at a given moment in time. Thus, this study used search volume data from the Google search engine to forecast the foreign tourist volume in the three target islands of Thailand.

For keywords searched related to tourism business, the Ministry of Tourism and Sports, Thailand (www.mots.go.th (2023)) has defined six relevant categories, namely accommodation, food and beverage, transportation, tour business, souvenir, and recreation. Each category is associated with the keywords shown in Table 2. These keywords were used to collect data on tourism-related keyword searches, while the data on arrivals in Thailand are described in the data preparation section below.

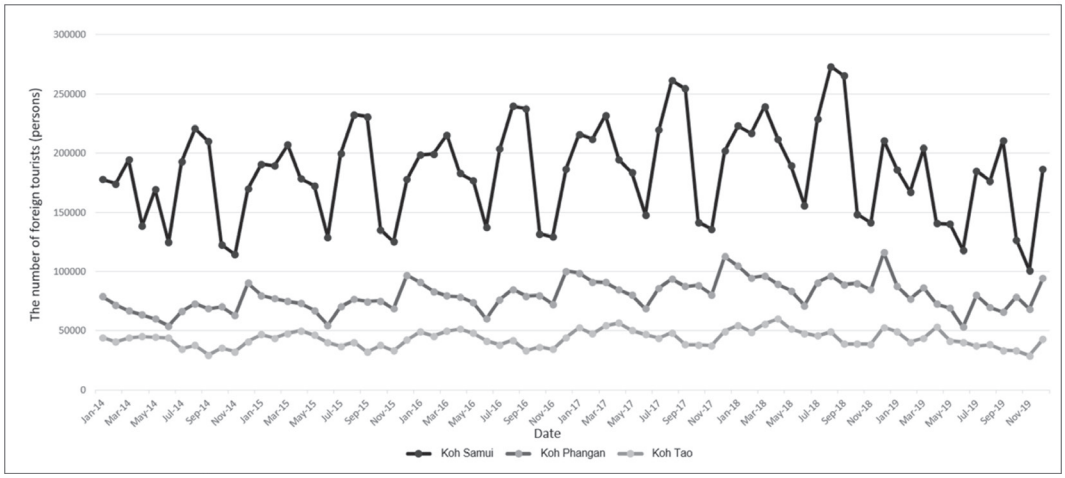
Table 2
The sets of words in the tourism business categories are provided by the Ministry of Tourism and Sports, Thailand

Accommodation	Food and beverage	Transportation	Tour business	Souvenir	Recreation
Hotel, Resort, Hostel, Inn, Guesthouse, Lodge, Hostelry, Model, Tavern, Pension, Boarding, Home stay, Farm stay, Bungalow, Cottage, Tourist - House, Condominium, Apartment, Dormitory, Villages, Campground, Caravan Parking	Fest food, Deli shop, Buffet, Coffee shop, Cafeteria, Gourmet, Night club, Restaurant, Eatery, Eating house, Seafood, Thai food restaurant, Grill restaurant, Ice cream shop, Juice shop, Sandwich shop, Hamburger shop, Hot dog shop, Vegetarian restaurant, Cafe, Bar	Bus, Taxi, Rental car, Charter coach Tour, Limousine, Train, Ferry, Flight	Package tour, Independent package tour, Hosted tour, Escorted tour, Affinity tour, Charter group, Group tour, Education tour, Religious tour, Scientific tour, Health tour, Sport tour, Ecotourism, Cultural tour, City tour, City-Sightseeing tour, Night tour, Trekking, Boat trip, Sea canoe, Diving, Cruise, Jungle raft, Safari, Half-Day tour, Day tour, Tour around, Guide	Shopping, Souvenir	Shopping center, Emporium, Market, Bistro, Diner, Restaurant, Canteen, Social center, Theme park, Amusement park, Museum, Garden, Park, Theater, Stadium, Fitness, Accommodation, Farm, Entertainment, Night club, Cafe, Bar, Pub, Casino

For the number of tourists in Thailand, we collect the data from the Ministry of Tourism and Sports website of Thailand (www.mots.go.th (2023)) for the years from 2014 to 2019 (inclusive). This study focuses only on foreign tourist arrivals in Thailand and only on English keywords in online searches. The study area comprises the three islands of Koh Samui, Koh Phangan, and Koh Tao in Surat Thani Province, Thailand, as this cluster of islands is among the top five tourist destinations in the country. The number of foreign tourist

arrivals by month from 2014 to 2019 is plotted in Figure 1. (The COVID-19 pandemic effectively started at the beginning of the year 2020, so this analysis pertains to the pre-pandemic situation.)

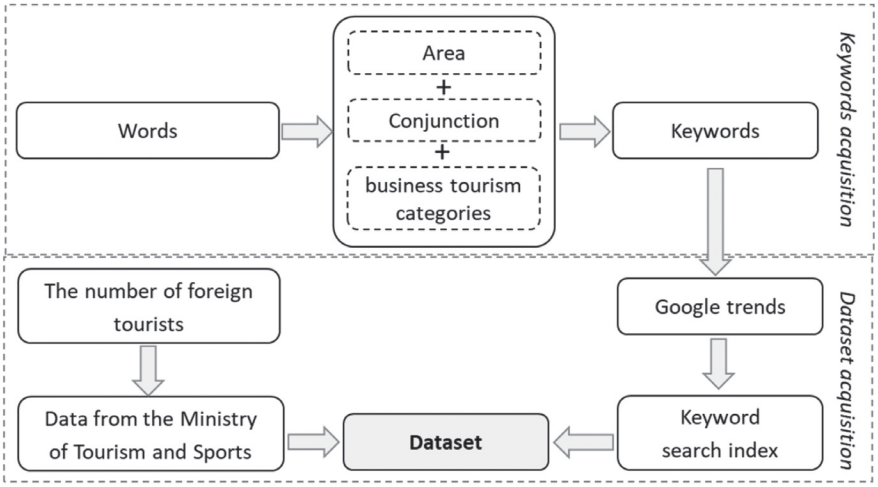
Figure 1
The monthly number of tourists arriving at three islands



3.2. Data preparation

Data preparation is a crucial step in this study, as illustrated in the flowchart in Figure 2, which outlines the steps involved in data preparation. To filter with the set of keywords, the first step is to define the possible terms for the search, including the study target area, link words, and business tourism categories. The targeted tourist attractions in this study were Koh Samui, Koh Phagan, and Koh Tao in Surat Thani province, Thailand. The conjunction or linking word is "in", indicating a relation to the study area with business tourism categories. The final term is the set of words from business tourism categories shown in Table 2. These three types of words can be permuted, or a keyword can be omitted; however, every keyword always has an accommodation and a location because the accommodation to search must be specified along with a specific area, which can be swapped.

Figure 2
The data preparation



To combine the keywords and obtain the dataset, the keywords were searched in Google Trends, and then the keyword search index for tourists was retrieved. Moreover, the relevant keywords were retrieved, along with their corresponding quantities. The final step is to combine the keyword index data and the number of foreign tourists into a three-type primary column dataset: time (month/year), the number of foreign tourists (in persons), and the keyword search index (volume). The keyword search index depends on the study area, and the datasets for Koh Samui, Koh Phagan, and Koh Tao consist of 334, 194, and 178 keywords, respectively. The dataset was collected and modelled to forecast the keyword search volumes and the amounts of tourist arrivals.

To test for causality, we applied the Granger Causality test using the statsmodels library in Python by importing the grangercausalitytests module. This test helps us determine if past values of the Google Trends keywords can significantly predict future tourist arrivals. For each keyword, we ran the test across different time lags, ranging from 1 to 5 months.

The Granger Causality test showed p-values for each keyword at different time lags. A p-value less than 0.05 indicates that the keyword search index for that particular Google Trends term significantly helps predict the number of tourist arrivals at that lag. For example: "Koh Samui Hotel" showed significant p-values at multiple lags, indicating a strong predictive relationship. Similarly, other keywords such as "TripAdvisor" and "us hostel koh samui" also showed significant results at certain lags.

These findings suggest that specific search terms on Google Trends are valid predictors of tourist arrivals. Therefore, incorporating keyword data from Google Trends in forecasting models could improve the accuracy of predictions related to tourism demand.

4. Research methodology

4.1. Framework of the methodology

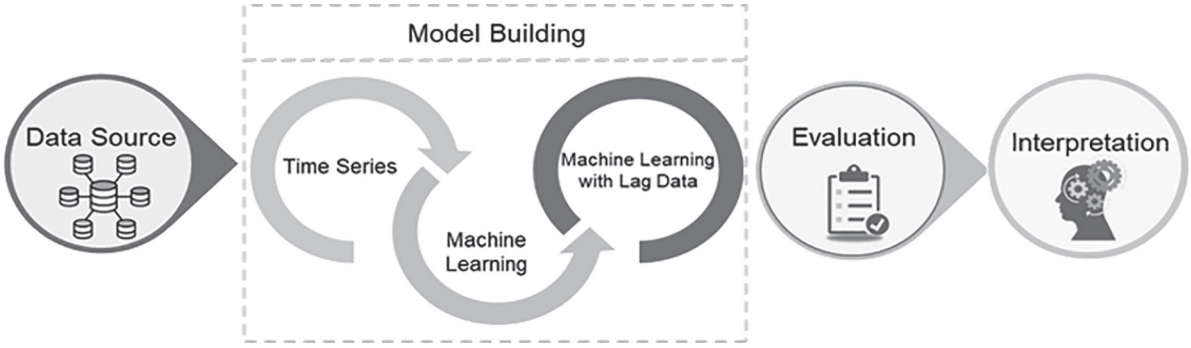
This section describes the overall methodological framework of the study, illustrated in Figure 3. Firstly, we prepared the source data as described in sections 3.1 and 3.2. The results from data preparation include the set of keywords and their usage counts from Google searches, as well as the number of foreign tourist arrivals each month.

Secondly, we applied machine learning methods to forecast foreign tourist demand from the search keywords as data sources. This work presents three types of models tested: time series, machine learning, and improved models utilising lag time data in machine learning. In the first studied model, a time series is applied to forecast the tourist arrivals from keyword search data. The advantage of this approach is that the forecasting is based on an interval of time. However, the set of keywords should relate to each other in terms of searches and be associated with real-time tourist arrivals. Thus, machine learning will also be studied in this work. In the machine learning part, we used ANN, Deep learning, Random Forest, and SVM approaches and compared their performances in forecasting foreign tourist volume. We use them to assess the predictive relationship between keyword searches and the volume of tourist arrivals. We note that tourists searching online do not necessarily arrive on the same day. Thus, we next improve the forecasting by adjusting for lag before building a model.

Thirdly, the results from trained (fitted) models are evaluated. The evaluation of model accuracy is a key part of the modelling process, assessing how well the model performs in forecasting. In this study, we choose Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) as the quantitative measures. These are commonly used in reporting prediction error rates in the performance of machine learning models.

Finally, the results of this study are discussed. We describe the output and/or outcome from data analysis using machine learning to show trends in forecasts of foreign tourist demand. More details are given in the next section.

Figure 3
An overview of the methodology used

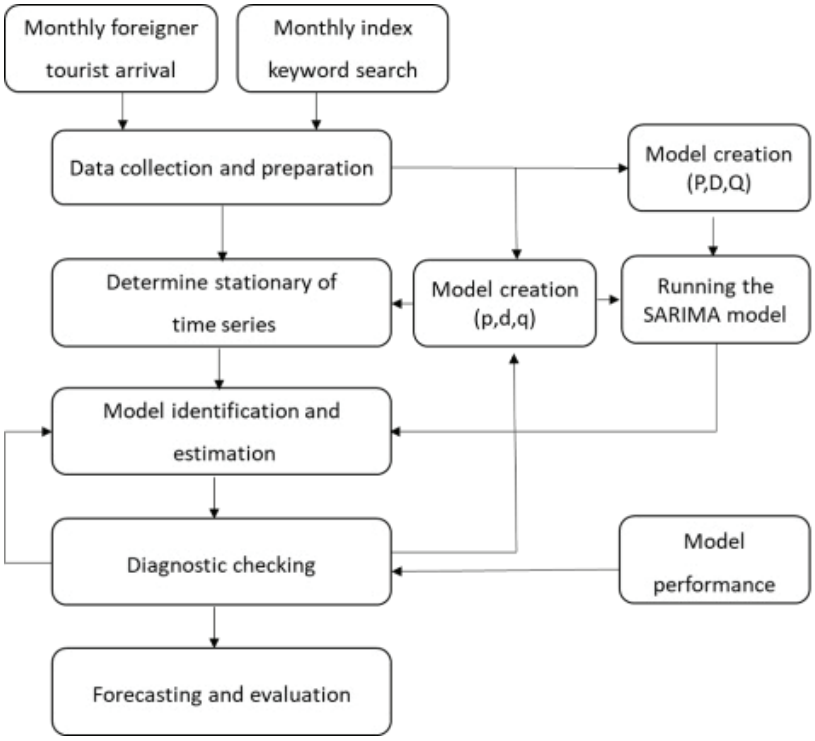


4.2. The time series model identification

The time series model is applied to forecast the volume of index keyword searches and the number of foreign tourist arrivals, as shown in Figure 4. The workflow in this figure begins with data collection and preparation, which involves extracting data from search engines and government organisations in a monthly format for the years 2014 to 2019. In this step, Pearson correlations are used to select relevant keywords associated with the number of tourist arrivals.

The obtained data will fit with the SARIMA model. It essentially combines the processes of different time series datasets using a combination of non-seasonal autoregressive (AR) and moving average (MA) models, along with the seasonal effects by period (S). The data set is used to find the model with configuration parameters (*p, d, q*) and (*P, D, Q*)s, where *p* is non-seasonal AR order, *d* is non-seasonal differencing, *q* is non-seasonal MA order, *P* is seasonal AR order, *D* is seasonal differencing, *Q* is seasonal MA order, and S is the period of repeating seasonal pattern. The model identified with (*p, d, q*) is used in the next step to determine a stationary time series and run the SARIMA model. The model identification and estimation are accomplished. Finding the sequence of changes needed to generate a static time series and then finding suitable forecasting models was pursued. Converting the time series through differencing is an essential part of the SARIMA model assembly process. Plotting the initial run sequence of the data indicated a rising trend. After identifying a model, it is necessary to perform diagnostic checks and test forecasting to evaluate model accuracy. These are a key part of the modelling process: assessing how well the model performs in predicting. In this study, we chose Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) as the measures.

Figure 4
The modelling steps with time series

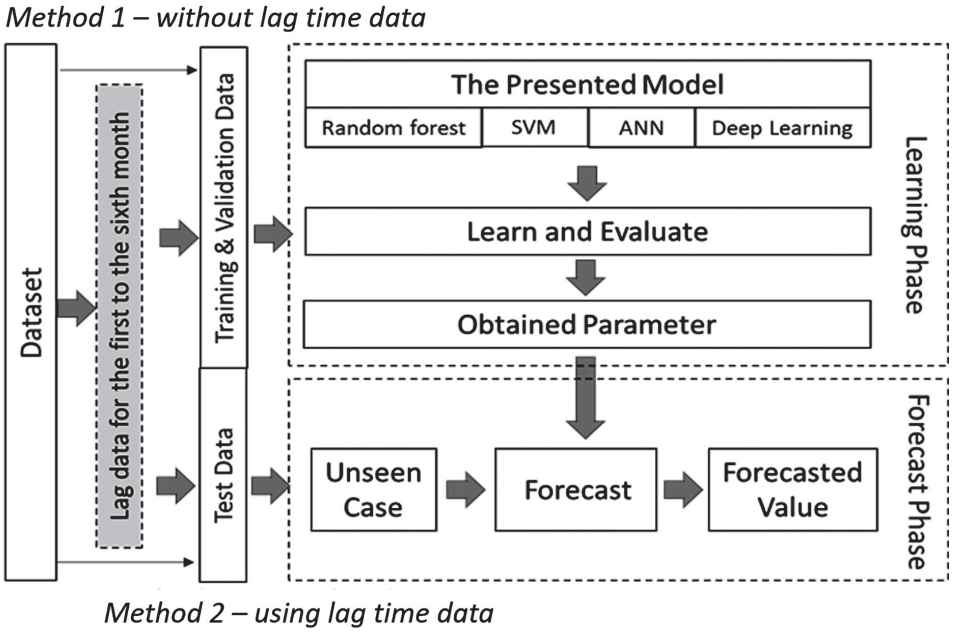


4.3. The comparison of alternative machine learning model types

Machine learning algorithms create, identify, and utilise trained models learned from data to predict future events. This process begins by analysing the information obtained and training (i.e., fitting model parameters) during the learning stage of the algorithms to create inference functions that can later predict model outputs for new input data. The system can easily provide targets for new inputs because there is sufficient training data. The learning algorithms compare model results with the target values to quantify errors and adjust model parameters, thereby reducing mistakes. Machine Learning allows the use of large data sets, while traditional statistics has focused on especially low-dimensional input data. ML usually provides the fastest and most accurate results, but it requires time and resources to train the predictive models.

This study applied machine learning to forecast foreign tourist arrivals using keyword search data, employing four types of algorithms: ANN, deep learning, Random Forest, and SVM. These algorithms were compared in terms of performance to select the best algorithm for predicting foreign tourist arrivals. The approach to modelling with machine learning methods is described in Figure 5, which consists of two stages in this study: one without lag time data and the other using lag time data.

Figure 5
The modelling steps with machine learning



The first method in Figure 5 is based on a data source without lag time. Namely, we prepare and use keyword searches to relate them with the number of tourist arrivals directly at each matching time. Afterwards, this dataset will be divided into two parts: training and test data. The training dataset will be used to learn model parameters for four types of algorithms: ANN, deep learning, random forest, and SVM. The training data are used to understand and evaluate the presented model, which achieves forecasting functions in a parametric form. These functions are then used to predict unseen cases in a separate test dataset, assessing the performance of the learned model on new data rather than its ability to fit the training data.

The second method in Figure 5 uses lag time data. When tourists search for information for travel planning, they do not depart immediately after the search. Instead, they leave on a later date. Thus, this work adjusts the dataset with lag time data before building a model. However, no data shows how many days after searching each tourist travel. For this reason, we adjust the relationship between keyword searches and the number of tourist arrivals, considering a time lag of zero to six months. Afterwards, the adjusted data are split into two groups, similarly as above, for training and a held-out dataset for testing.

To evaluate the performance of the proposed model using machine learning, the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are employed, following the same approach as with the time series model. The results are summarized in the next section.

5. Results and discussion

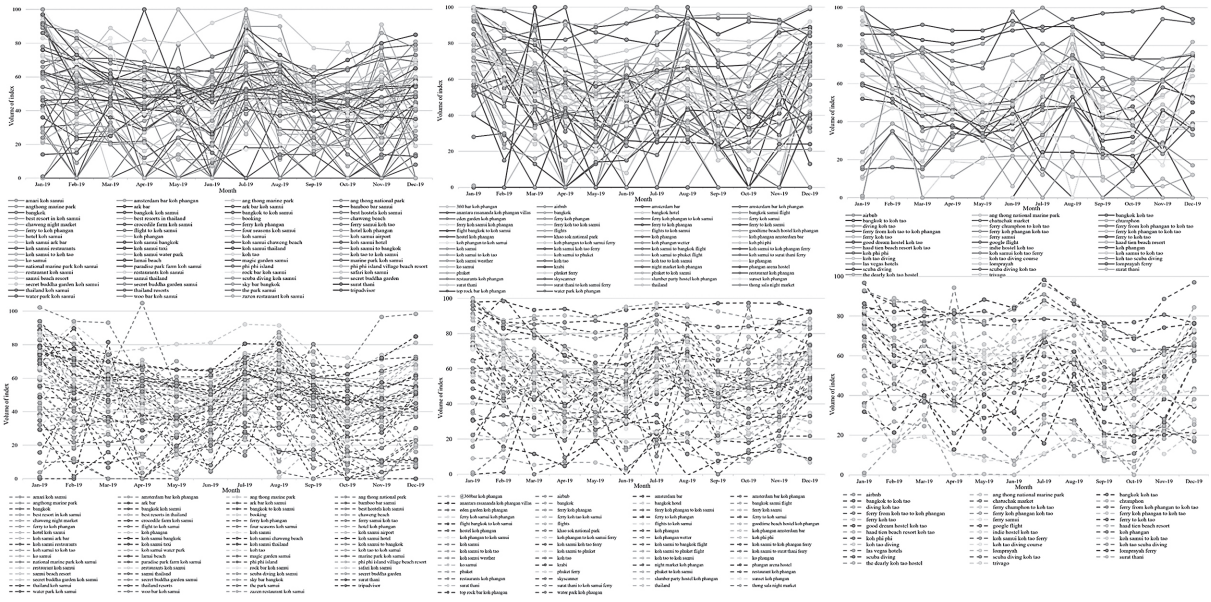
5.1. Time series performance

The data obtained, described in sections 3.1 and 3.2, consists of keyword searches and the number of tourists from 2014 to 2019, used to identify predictive relationships. The Pearson correlations and their significance are used to assess the relevance of tourist arrivals. The set of keywords associated with traveller arrivals is used, while some irrelevant keywords are eliminated — in this step, we perform feature selection. Afterwards, the time series SARIMA model is trained using data from the years 2014 to 2018. The remaining year, 2019, is set aside for testing the trained model in forecasting tourist arrivals. We fit the model (p, d, q) appropriately for each keyword (23 models), with two models shown as examples in Table 3. The forecasting of tourism-related keyword searches is illustrated in Figure 6, where the upper figures display actual data, and the lower figures show model outputs for Koh Samui, Koh Phangan, and Koh Tao.

Table 3
Examples of the fit SARIMA model orders for each keyword

Model (p,d,q)	Koh Samui	Koh Phangan	Koh Tao
(0,0,0)	Ang thong marine park Best hostels Koh Samui Best resort in Koh Samui Crocodile farm Koh Samui Paradise park farm Koh Samui Secret buddha garden Koh Samui Secret buddha garden Samui Thailand resorts Woo bar Koh Samui Zazen restaurant Koh Samui	Accommodation in Koh Phangan Bangkok samui flight Eden garden Koh Phangan Ferry Koh Phangan to Koh Samui Flight bangkok to Koh Samui Koh Phangan to Koh Samui ferry Koh Samui to bangkok Koh Samui to bangkok flight Koh Samui to Koh Phangan ferry Koh Samui to phuket flight Koh Samui to surat thani ferry Surat thani to Koh Samui ferry thailand Water park Koh Phangan	Ang thong national marine park Ferry chumphon to Koh Tao Ferry from Koh Phangan to Koh Tao Ferry from Koh Tao to Koh Phangan Koh Tao scuba diving Ocean view bungalows Koh Tao Scuba diving Koh Tao
…	…	…	…
(0,1,1)	Amsterdam bar Koh Phangan Bangkok Bangkok to Koh Samui Best resorts in thailand Ferry Samui Koh Tao Ferry to Koh Phangan Koh Samui taxi Koh Samui to bangkok Koh Samui to Koh Tao Koh Tao to Koh Samui Marine park Koh Samui National marine park Koh Samui Safari Koh Samui Scuba diving Koh Samui The park samui	Anantara rasananda Koh Phangan villa resort & spa Ferry Koh Phangan Ferry Koh Tao Koh Samui Goodtime beach hostel Koh Phangan Hostel Koh Phangan Khao sok national park Koh Samui Koh Tao ferry Santhiya Koh Phangan resort & spa Santhiya resort & spa Koh Phangan Skyscanner	Chatuchak market Diving Koh Tao Ferry to Koh Tao Good dream hostel Koh Tao Google flight Indie hostel Koh Tao Koh Phangan Koh Tao diving Spice market The dearly Koh Tao hostel
…	…	…	…

Figure 6
Forecast and actual keyword searches using time series models



In Figure 6, the results are difficult to interpret in terms of trends in keyword searches. Thus, to easily describe and illustrate trends from forecasting, an example of keyword searching for each study area is provided in Figure 7 for four keywords associated with the Koh Samui, Koh Phangan, and Koh Tao areas. Each graph shows similar trends in the forecast and observed data. Moreover, the trend in number of tourist arrivals for each study area in Figure 8 has a similar seasonal pattern. Namely, the forecast results and actual data are concordant in the trend direction. Therefore, the model appears effective.

Figure 7
Example forecast and actual keyword search from Figure 6

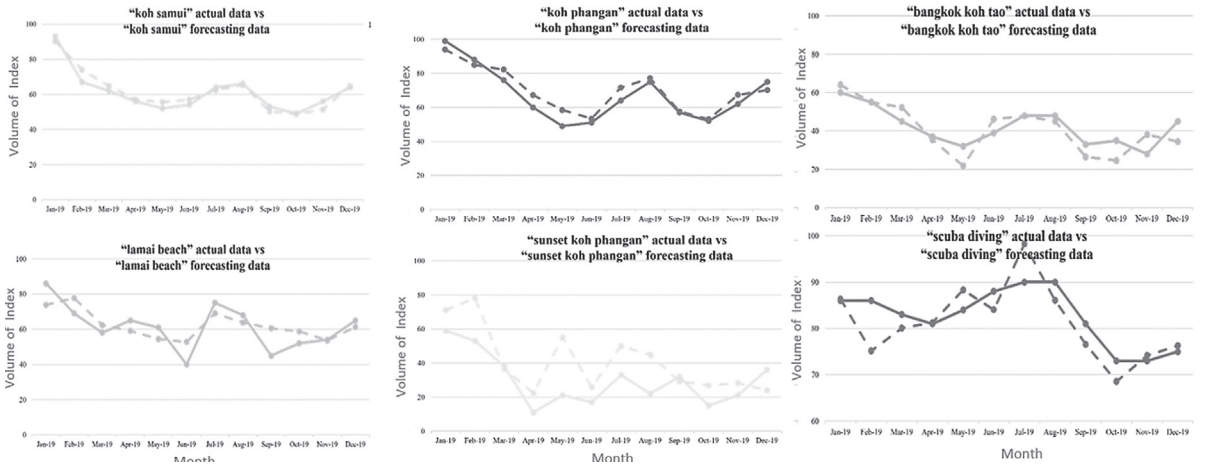
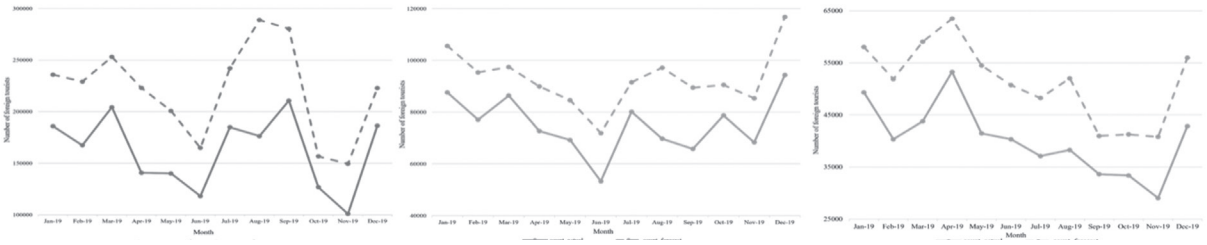


Figure 8
The forecast and actual tourist arrivals using a time series model



The example in Figure 7 demonstrates the performance of forecasting keyword searches and the number of tourist arrivals. MAE, MSE and RMSE were used to evaluate the models. The results of this study, in terms of the model performance measures MAE, MSE, and RMSE, are shown in Table 4. It was found that the trend in search index volume of selected keywords was similar in pattern to the number of tourist arrivals.

Table 4
The performance evaluation of the time series model is from Figure 7

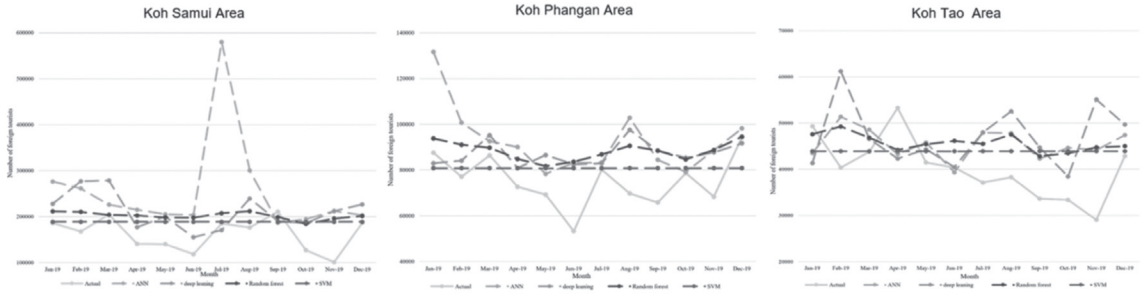
Area	Keyword	MAE	MSE	RMSE
Koh Samui	Koh Samui	2.75	11.08	3.33
	Lamai beach	7.33	72.33	8.50
Koh Phangan	Koh Phangan	4.50	27.00	5.20
	Sunset Koh Phangan	13.84	274.34	16.57
Koh Tao	Bangkok Koh Tao	5.83	48.67	6.98
	Scuba diving	3.67	23.00	4.80

The SARIMA model is effective in forecasting because the actual data and the forecast data tend to go in the same direction for each keyword, having similar seasonal patterns. Likewise, the forecast number of tourist arrivals tends to change in the same direction as the actual data, satisfying forecasting efficiency. However, this model does not capture the relationships between search activity and real-time tourist arrivals. To address this limitation, machine learning will now be applied to identify the factors that influence foreign tourist volume.

5.2. Performance comparison of machine learning models and of using lag time

Machine learning was applied to predict the number of tourist arrivals from a set of keyword searches. This study divides the dataset into two parts for training and testing. The training data spanned years from 2014 to 2018, while the data for 2019 were held out as a test dataset for forecasting tourist arrivals. In the training phase, 10-fold cross-validation was used. A comparison between the four algorithms — ANN, deep learning, Random Forest, and SVM — for the number of tourist arrivals in the three study areas (Koh Samui, Koh Phangan, and Koh Tao) is shown in Figure 9.

Figure 9
Comparing tourist arrival predictions from machine learning models for each area



From Figure 9, we found that the SVM forecast was closest to actual data in 2019. To assess the performance in predicting tourist arrivals, MAE, MSE, and RMSE were used to evaluate the models. The results are presented in Table 5, with the last column. We found that the SVM algorithm had the best performance among the tested algorithms. However, when using data from 2014-2018 for training and testing with a one-year-out cross-validation, the Random Forest algorithm appears to have the best performance, as shown in columns 4-8 of Table 5.

Table 5
The forecasting performance of machine learning models in each year

Area	Performance evaluation	Models	Test Data in Year					
			2014	2015	2016	2017	2018	2019
Koh Samui	MAE	ANN	39471.59	29364.97	35349.91	37553.62	45400.13	97379.01
		Deep learning	42797.05	35255.89	34160.82	48963.92	65340.00	44346.54
		Random forest	**28864.24**	**25091.14**	**23470.51**	**30140.91**	**40289.15**	42022.00
		SVM	28864.24	27627.88	29578.81	38616.38	44703.70	**33044.35**
	MSE	ANN	2996997545.76	1433693080.52	1720496493.08	1878087622.02	2924769317.25	18622617035.25
		Deep learning	2554666463.11	1884135060.40	1589305621.12	3811404338.86	5389643940.44	2660995183.04
		Random forest	**1214009301.66**	**1052085077.05**	**979208836.63**	**1388683932.86**	**2327838608.88**	2528100483.65
		SVM	1214009301.66	1208305489.41	1312800836.61	1949356122.25	2551115123.54	**1877632921.39**
	RMSE	ANN	54744.84	37864.14	41478.87	43336.91	54081.14	136464.62
		Deep learning	50543.71	43406.62	39866.10	61736.57	73414.19	51584.83
		Random forest	**34842.64**	**32435.86**	**31292.31**	**37265.05**	**48247.68**	50280.22
		SVM	34842.64	34760.69	36232.59	44151.51	50508.56	**43331.66**
Koh Phangan	MAE	ANN	7549.05	9388.49	12094.22	12656.99	18329.59	17481.62
		Deep learning	12031.04	8347.58	7591.43	9293.05	13024.40	18741.70
		Random forest	**9537.53**	**5886.01**	**6142.21**	**8087.04**	**13320.43**	12941.15
		SVM	14500.59	9496.63	6591.52	12066.66	15874.10	**9834.81**
	MSE	ANN	90838574.66	112600579.51	226604856.31	339106312.69	466563177.84	474573349.44
		Deep learning	193946466.02	102434081.18	93653032.16	137122024.21	257082040.20	437184720.68
		Random forest	**106446769.75**	**65323193.30**	**72432392.22**	**109193478.99**	**239150553.81**	242589492.61
		SVM	251742208.01	134025453.65	90316916.92	216767572.70	333595888.34	**144640869.41**
	RMSE	ANN	9530.93	10611.34	15053.40	18414.84	21600.07	21784.70
		Deep learning	13926.47	10120.97	9677.45	11709.91	16033.78	20908.96
		Random forest	**10317.30**	**8082.28**	**8510.72**	**10449.57**	**15464.49**	15575.28
		SVM	15866.39	11576.94	9503.52	14723.03	18264.61	**12026.67**
Koh Tao	MAE	ANN	7785.57	4413.40	8187.09	4854.84	5861.86	8076.54
		Deep learning	7687.34	5344.07	8007.73	4647.69	6218.81	7260.85
		Random forest	**4686.62**	**4746.90**	**4923.62**	**4765.86**	**6778.22**	7270.40
		SVM	4900.35	4933.10	5214.47	6759.95	8095.26	**6129.63**
	MSE	ANN	127468848.12	30612670.95	97067093.53	46241553.77	60504362.52	79718594.72
		Deep learning	101479454.43	41391935.17	85827767.87	38949479.47	78924076.24	84062158.47
		Random forest	**36020808.22**	**30732664.89**	**33949287.48**	**34530679.13**	**73786308.08**	67983377.41
		SVM	47351771.82	35211266.53	36072027.66	57990869.14	86472687.15	**55323653.64**
	RMSE	ANN	11290.21	5532.87	9852.26	6800.11	7778.46	8928.53
		Deep learning	10073.70	6433.66	9264.33	6240.95	8883.92	9168.54
		Random forest	**6001.73**	**5543.70**	**5826.60**	**5876.28**	**8589.90**	8245.20
		SVM	6881.26	5933.91	6006.00	7615.17	9299.07	**7437.99**

An interesting issue is why the model changed in 2019 for predicting tourist arrivals in the study area. According to Figure 1, which shows the tourist arrivals in 2019, the number of tourists began to decline. As potential

reasons, there was the strengthening of the Thai currency (baht), the US-China trade war, and uncertainty regarding the European Union's relationship with the United Kingdom. These might be the reasons why the number of tourists decreased. The tourism business in the study area has become quite challenging with the ongoing epidemic, as more than 80% of the tourists in these three areas were foreigners.

Additionally, the machine learning models support predicting tourist arrivals based on keyword search data. This also highlights changes that have occurred, which could lead to a root cause analysis and inform guidelines for resolving further problems.

Considering the keyword searches, tourism has a lag time, and the keyword searches do not co-occur with actual travel. Thus, time lags from zero to six months were tested. According to the above results, the Random Forest model appears to be the best choice for predicting tourist arrivals from 2014 to 2018. For this reason, this model is used to indicate the number of tourists from search data, taking into account lag times. The correlation between actual and predicted values with lag times in the years 2014-2018 was assessed to determine the time lag between online searches and actual visits in the study areas of Koh Samui, Koh Phangan, and Koh Tao. We used two types of correlations, namely Pearson and Spearman, as shown in Table 6.

Table 6
Correlations of model outputs when using time lags for each study area

Area	Correlation method	Lag time from the first to sixth month					
		Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6
Koh Samui	Pearson	0.218	0.460	**0.465**	0.162	0.302	0.302
	Spearman	0.241	0.446	**0.466**	0.182	0.269	0.269
Koh Phangan	Pearson	0.620	0.460	0.491	0.519	**0.645**	0.582
	Spearman	0.566	0.446	0.503	0.536	**0.649**	0.593
Koh Tao	Pearson	0.517	0.581	**0.660**	0.679	0.43	0.467
	Spearman	0.544	0.547	**0.669**	0.644	0.38	0.465

In Table 6, the results indicate that the time lag after searching with keywords varies depending on the target area to be visited. Namely, for the Koh Samui and Koh Tao areas, a 3-month lag showed significant correlations, while for the Koh Phangan area, a lag of 5 months yielded the highest correlations.

Based on the above results, we identified the top 10 most correlated keywords for each area using these lag times, which are presented in Table 7. These keywords indicate that the demands of foreign tourists in the three study areas of interest are similar, encompassing accommodation, booking websites, activities, attractions, and transportation. For the benefit of entrepreneurs in the three study areas, the demands of tourists were analysed separately by area for keyword reach, business tourism industry, and behavioural segmentation. The results correspond to the business tourism industry of Thailand, which includes accommodation, food and beverages, transportation, tour businesses, souvenirs, and recreation. According to the behaviour of the activities of foreign tourists, the Tourism Authority of Thailand has divided foreign tourists into five groups: (1) The "Eat.Play.Shop", (2) The Activities Explorer, (3) The Beach & Night Wanderer, (4) The City Nomads, and (5) The Ocean Lover (Tourism Authority of Thailand (2020)). To analyse tourist demands, the set of obtained keywords was used to assess business tourism industry categories and activities (tourist targets), with the results presented in Table 8.

This research will support tourism operators and stakeholders in planning how to cater to future foreign tourists and effectively meet their demands. Moreover, the set of obtained keywords can be utilised in Search Engine Optimisation (SEO) to enhance visibility in tourist searches.

Table 7
The top 10 most correlated keywords

Koh Samui	Koh Phangan	Koh Tao
Chatuchak market	Airbnb	Scuba diving
Koh Samui bungalow am strand	Booking	Booking
Best bars in Koh Samui	Skyscanner	Trivago
Ark bar Ko Samui	Flights	Dream hotel
Hostels Koh Samui	Ang thong national marine park	Airbnb
Calypso diving Koh Samui	Amsterdam bar	Hostel world
Best hostels Koh Samui	Three sixty bar Koh Phangan	Ang thong national marine park
Ferry Koh Phangan to Koh Samui	Trivago	Koh Tao bar crawl
Koh Phangan to Koh Samui ferry	360 bar Koh Phangan	Skyscanner
Koh Samui to Koh Phangan ferry	High life bungalow Koh Phangan	Bars near me

Table 8
Tourism demands are based on keyword searches for each area

Area	Keywords	Business tourism industry	Tourist target
Koh Samui	Chatuchak market	Souvenir, Reaction	The "Eat.Play.Shop"
	Koh Samui bungalow am strand	Accommodation	-
	Best bars in Koh Samui	Food and beverage, Recreation	The "Eat.Play.Shop", The beach & Night wanderer
	Ark bar Ko Samui	Food and beverage, Recreation	The "Eat.Play.Shop", The beach & Night wanderer
	Hostels Koh Samui	Accommodation	-
	Calypso diving Koh Samui	Recreation	Ocean lover
	Best hostels Koh Samui	Accommodation	-
	Ferry Koh Phangan to Koh Samui	Transportation	-
	Koh Phangan to Koh Samui ferry	Transportation	-
	Koh Samui to Koh Phangan ferry	Transportation	-
Koh Phangan	Airbnb	Tour business	-
	Booking	Tour business	-
	Skyscanner	Tour business	-
	Flights	Transportation	-
	Ang thong national marine park	Recreation	The "Eat.Play.Shop", Ocean lover
	Amsterdam bar	Food and beverage, Recreation	The "Eat.Play.Shop", The beach & Night wanderer
	Three sixty bar Koh Phangan	Food and beverage, Recreation	The "Eat.Play.Shop", The beach & Night wanderer
	Trivago	Tour business	-
	360 bar Koh Phangan	Food and beverage, Recreation	The "Eat.Play.Shop", The beach & Night wanderer
	High life bungalow Koh Phangan	Accommodation	-
Koh Tao	Scuba diving	Recreation	Ocean lover
	Booking	Tour business	-
	Trivago	Tour business	-
	Dream hotel	Accommodation	-
	Airbnb	Tour business	-
	Hostel world	Accommodation	-
	Ang thong national marine park	Recreation	The "Eat.Play.Shop", Ocean lover
	Koh Tao bar crawl	Food and beverage, Recreation	The "Eat.Play.Shop", The beach & Night wanderer
	Skyscanner	Tour business	-
	Bars near me	Food and beverage, Recreation	The "Eat.Play.Shop", The beach & Night wanderer

6. Conclusion

This study highlights the potential of using keyword search data from Google Trends to predict tourist arrivals. The study focused on the three islands of Koh Samui, Koh Phangan, and Koh Tao in Surat Thani Province, Thailand. These are among the top five most popular tourist destinations in the country. These areas were used to study the relationships between keyword searches by potential tourists and their subsequent arrival volumes. Thus, machine learning was applied to identify predictive relationships. The data on keyword searches were derived from Google Trends, while the data on tourist arrivals were obtained from the Ministry of Tourism and Sports of Thailand's website from January 2014 to December 2019. The MAE, MSE, and RMSE were used to evaluate the model's fit accuracy and prediction performance. On applying machine learning models, such as SARIMA, ANN, Random Forest, and SVM, the results demonstrate that keyword searches are closely linked to future tourism demand, where the time series approach using the SARIMA model can forecast the volume trends in both keyword searches and in tourist arrivals, these having similar seasonal patterns. These findings have significant implications for tourism operators and stakeholders, as they offer a data-driven approach to anticipating shifts in tourist interest and adjusting marketing and operational strategies accordingly.

On the other hand, this study gained knowledge from testing machine learning algorithms, including ANN, Deep Learning, Random Forest, and SVM. The most effective model, SVM, outperformed others in predicting tourist arrivals for the 2019 held-out test data, indicating its reliability in real-world forecasting. Additionally, the study revealed that the predictive power of keyword searches varies by time lag and location, suggesting that customised forecasting models should be considered for different regions. These insights enable tourism businesses to optimize resources, enhance visitor experiences, and improve strategic planning based on evolving tourist preferences. By leveraging search data, tourism operators can proactively address future demands, ensuring better alignment with market trends and improving their competitive edge in a rapidly changing industry.

Acknowledgements

The authors are very grateful to the Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Thailand. This research was supported by National Science, Research and Innovation Fund (NSRF) and Prince of Songkla University (Ref. No. SIT6701162S).

References

Adulwattana, B., & Pitakard, B. (2019, May 16) Tourism: Still a reliable driver of growth? *Bangkok Bank Research*. https://www.bangkokbank.com/th-TH/International-Banking/-/media/dc98bd4dd875455299b0c207bf16f2ac.ashx#:~:text=The%20number%20of%20visitors%20to,in%202009%20(Chart%202)

Alvarez-Díaz, M., González-Gómez, M., & Otero-Giráldez, M.S. (2019). Forecasting international tourism demand using a nonlinear autoregressive neural network and genetic programming. *Forecasting, 1*(1), 90–106. https://doi.org/10.3390/forecast1010007

Asaithambi, S.P.R., Venkatraman, R., & Venkatraman, S. (2023). A thematic travel recommendation system using an augmented big data analytical model. *Technologies, 11*(1), 28. https://doi.org/10.3390/technologies11010028

Ascher-Walsh, R. (2021, September 8). Our Readers' Favorite five islands in Asia in 2021. *Travel + Leisure*. https://www.travelandleisure.com/worlds-best/islands-in-asia

Bi, J.W., Liu, Y., & Li, H. (2020). Daily tourism volume forecasting for tourist attractions. *Annals of Tourism Research, 83,* Article 102923. https://doi.org/10.1016/j.annals.2020.102923

Chandra, S., & Kumari, K. (2018). Forecasting foreign tourist arrivals in India using time series models. *International Journal of Statistics and Applied Mathematics, 3*(2), 338-342. https://www.mathsjournal.com/archives/2018/vol3/issue2/PartE/3-1-81

Cheng, X., Fu, S., & De Vreede, G.J. (2018). A mixed method investigation of sharing economy driven car-hailing services: Online and offline perspectives. *International Journal of Information Management, 41*, 57-64. https://doi.org/10.1016/j.ijinfomgt.2018.03.005

Dinis, G., Breda, Z., Costa, C., & Pacheco, O. (2019). Google trends in tourism and hospitality research: a systematic literature review. *Journal of Hospitality and Tourism Technology, 10*(4), 747-763. https://doi.org/10.1108/JHTT-08-2018-0086

Feng, Y., Li, G., Sun, X., & Li, J. (2019). Forecasting the number of inbound tourists with Google Trends. *Procedia Computer Science, 162*, 628–633. https://doi.org/10.1016/j.procs.2019.12.032

Kaewmanee, P., Muangprathub, J., & Sae-jie, W. (2021). Forecasting tourist arrivals with keyword search using time series. In Y. Kumsuwan (Ed.), *ECTI-CON 2021 - 2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* (pp. 171–174). IEEE. https://doi.org/10.1109/ECTI-CON51831.2021.9454824

Khatibi, A., Belém, F., da Silva, A.P.C., Almeida, J.M., & Gonçalves, M.A. (2020). Fine-grained tourism prediction: Impact of social and environmental features. *Information Processing & Management, 57*(2), Article 102057. https://doi.org/10.1016/j.ipm.2019.102057

Law, R., Li, G., Fong, D.K.C., & Han, X. (2019). Tourism demand forecasting: A deep learning approach. *Annals of Tourism Research, 75*, 410-423. https://doi.org/10.1016/j.annals.2019.01.014

Li, H., Hu, M., & Li, G. (2020). Forecasting tourism demand with multisource big data. *Annals of Tourism Research, 83*, Article 102912. https://doi.org/10.1016/j.annals.2020.102912

Li, K., Lu, W., Liang, C., & Wang, B. (2019). Intelligence in tourism management: A hybrid FOA-BP method on daily tourism demand forecasting with web search data. *Mathematics, 7*(6),1-14.

Li, X., Law, R., Xie, G., & Wang, S. (2021). Review of tourism forecasting research with internet data. *Tourism Management, 83*, Article 104245. https://doi.org/10.1016/j.tourman.2020.104245

Liu, A., & Wu, D.C. (2019). Tourism productivity and economic growth. *Annals of Tourism Research, 76*, 253-265. https://doi.org/10.1016/j.annals.2019.04.005

Pan, B., Chenguang Wu, D., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology, 3*(3), 196–210. https://doi.org/10.1108/17579881211264486

Peng, L., Wang, L., Ai, X.-Y., & Zeng, Y.-R. (2020). Forecasting Tourist Arrivals via Random Forest and Long Short-term Memory. *Cognitive Computation, 13*(1), 125–138. https://doi.org/10.1007/s12559-020-09747-z

Sachdev, C. (2021, September 8). Our readers' favorite 10 Southeast Asia resort hotels in 2021. *Travel + Leisure*. https://www.travelandleisure.com/worlds-best/resort-hotels-in-southeast-asia

Sun, S., Wei, Y., Tsui, K.-L., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management, 70*, 1–10. https://doi.org/10.1016/j.tourman.2018.07.010

Wen, L., Liu, C., & Song, H. (2019). Forecasting tourism demand using search query data: A hybrid modelling approach. *Tourism Economics, 25*(3), 309-329. https://doi.org/10.1177/1354816618768317

Wen, L., Liu, C., Song, H., & Liu, H. (2021). Forecasting tourism demand with an improved mixed data sampling model. *Journal of Travel Research, 60*(2), 336-353. https://doi.org/10.1177/0047287520906220

Wickramasinghe, K., & Ratnasiri, S. (2020). The role of disaggregated search data in improving tourism forecasts: Evidence from Sri Lanka. *Current Issues in Tourism, 24*(19), 2740–2754. https://doi.org/10.1080/13683500.2020.1849049

Wongsathan, R. (2018). SARIMA intervention based forecast model for visitor arrivals to Chiang Mai, Thailand. *Asia-Pacific Journal of Science and Technology, 23*(4), 1-14. https://doi.org/10.14456/apst.2018.19

Wu, D.C.W., Ji, L., He, K., & Tso, K.F.G. (2021). Forecasting tourist daily arrivals with a hybrid Sarima–Lstm approach. *Journal of Hospitality & Tourism Research, 45*(1), 52-67. https://doi.org/10.1177/1096348020934046

Xie, G., Qian, Y., & Wang, S. (2021). Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach. *Tourism Management, 82*, Article 104208. https://doi.org/10.1016/j.tourman.2020.104208

Yaghini, M., Khoshraftar, M.M., & Fallahi, M. (2013). A hybrid algorithm for artificial neural network training. *Engineering Applications of Artificial Intelligence, 26*(1), 293-301. https://doi.org/10.1016/j.engappai.2012.01.023

Yang, Y., Guo, J., & Sun, S. (2021). Tourism demand forecasting and tourists' search behavior: Evidence from segmented Baidu search volume. *Data Science and Management, 4*, 1-9. https://doi.org/10.1016/j.dsm.2021.10.002

Zervas, G., Proserpio, D., & Byers, J.W. (2017). The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry. *Journal of Marketing Research, 54*(5), 687-705. https://doi.org/10.1509/jmr.15.0204

Submitted: July 19, 2024
Revised: December 26, 2024
Accepted: January 10, 2025