

# IMPACT OF EXTERNAL FACTORS ON AIR PASSENGER DEMAND PREDICTION USING MACHINE LEARNING

Miss Sutthiya Lertyongphati



A Thesis Submitted in Partial Fulfillment of the Requirements  
for the Degree of Master of Science in Computer Science and  
Information Technology  
Department of Mathematics and Computer Science  
Faculty Of Science  
Chulalongkorn University  
Academic Year 2023

ผลกระทบจากปัจจัยภายนอกต่อความต้องการผู้โดยสารสายการบิน โดยการเรียนรู้ของเครื่อง



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต  
สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ภาควิชาคณิตศาสตร์และวิทยาการ  
คอมพิวเตอร์  
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย  
ปีการศึกษา 2566

Thesis Title	IMPACT OF EXTERNAL FACTORS ON AIR PASSENGER DEMAND PREDICTION USING MACHINE LEARNING
By	Miss Sutthiya Lertyongphati
Field of Study	Computer Science and Information Technology
Thesis Advisor	Assistant Professor Dr. DITTAYA WANVARIE

---

Accepted by the FACULTY OF SCIENCE, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

..... Dean of the FACULTY OF SCIENCE  
(Professor Dr. PRANUT POTIYARAJ)

THESIS COMMITTEE

..... Chairman  
(Associate Professor Dr. NAGUL COOHAROJANANONE)

..... Thesis Advisor  
(Assistant Professor Dr. DITTAYA WANVARIE)

..... External Examiner  
(Dr. Ukrit Watchareeruetai)

  
จุฬาลงกรณ์มหาวิทยาลัย  
CHULALONGKORN UNIVERSITY

สุทธิญาณ์ เลิศขงผาดิ : ผลกระทบจากปัจจัยภายนอกต่อความต้องการผู้โดยสารสายการบินโดยใช้การเรียนรู้ของเครื่อง. ( IMPACT OF EXTERNAL FACTORS ON AIR PASSENGER DEMAND PREDICTION USING MACHINE LEARNING) อ.ที่ปรึกษาหลัก : ผศ. ดร.ทิตขา หวานวาริ

ปัจจัยภายนอกสามารถส่งผลกระทบต่อปริมาณผู้โดยสารที่เดินทางโดยเครื่องบินหรือไม่งานวิจัยนี้นำเสนอการหาความสัมพันธ์ระหว่างปัจจัยภายนอกที่ส่งผลกระทบต่อปริมาณผู้โดยสารโดยเครื่องบินโดยใช้การเรียนรู้ของเครื่องสามวิธี ได้แก่ Gradient Boosting, Random Forest, and Support Vector Regression โดยใช้ข้อมูลปริมาณผู้โดยสารขาเข้าประเทศโดยเครื่องบิน ระหว่างเดือนมกราคม พ.ศ. 2554 ถึงธันวาคม พ.ศ. 2562 แบ่งออกตามภูมิภาค ดังต่อไปนี้ 1) ภาคกลาง 2) ภาคเหนือ 3) ภาคใต้ รวมถึงข้อมูลปัจจัยภายนอกต่างๆ ซึ่งประกอบด้วย 1) ข้อมูลการค้นหาค่าที่เกี่ยวข้องกับปริมาณผู้โดยสารที่เดินทางเข้าประเทศจาก Google Trend 2) ข้อมูลสภาพภูมิอากาศ 3) จำนวนงานคอนเสิร์ตและการแข่งขันกีฬา 4) ดัชนีราคาผู้บริโภค และราคาน้ำมันเชื้อเพลิงอากาศยาน

ปัจจัยทั้งหมดถูกนำมาสร้างโมเดลในการทำนายปริมาณผู้โดยสารขาเข้าโดยเครื่องบิน แบ่งตามภูมิภาคดังที่กล่าวข้างต้น ผลลัพธ์จากการทำนายที่ประกอบด้วยปัจจัยภายนอกนั้นมีความคลาดเคลื่อน RMSE ต่ำกว่าการทำนายโดยการใช้ข้อมูลผู้โดยสารขาเข้าเพียงอย่างเดียว ซึ่งให้เห็นว่าข้อมูลสภาพอากาศประกอบค่าที่ใช้การค้นหาค่าจาก Google Trend รวมถึงดัชนีราคาผู้บริโภคและราคาน้ำมันเชื้อเพลิงอากาศยาน ต่างส่งผลกระทบต่อปริมาณผู้โดยสารขาเข้ามากที่สุด แต่ลักษณะอากาศ รวมถึงค่าที่ใช้การค้นหาค่าจาก Google Trend ที่ส่งผลกระทบนั้นต่างออกไปในแต่ละภูมิภาค สรุปได้ว่าปัจจัยภายนอกที่เกี่ยวข้องสามารถนำมาใช้เพื่อประกอบการพยากรณ์หรือคาดการณ์แนวโน้มปริมาณผู้โดยสารที่เดินทางโดยเครื่องบินได้อย่างมีประสิทธิภาพ



สาขาวิชา	วิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ	ลายมือชื่อนิสิต .....
ปีการศึกษา	2566	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

## 6172628823 : MAJOR COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

KEYWORD air travel demand, machine learning

:

Do external factors impact the volume of air passengers? Can they be employed to analyze the dependency of hidden parameters in demand forecast strategies? In this research, a framework is proposed to investigate the impact of external factors on demand for air travel by combining the features extracted from various platforms with the historical volume of inbound passenger data from January 2011 to December 2021 and comparing the information symmetry to uncover relations between the data, proving whether they contributed to the shift in demand. A selection of machine learning regression models, namely, gradient boosting, random forest, and support vector regression, were utilized to build a prediction model with and without the inclusion of the additional variables. Their performance will justify our assumption of the impact of external factors on passengers traveling by air. Employing Thailand's historical inbound passenger volume, the result had shown that with the addition of explanatory variables had reduced RMSE. A combination of certain weather elements and search queries has the most impact on the air travel demand in Thailand, but the combination varies in each region. Event indicators and econometric variables introduce further enhancement in accordance with the preliminary assumption of their influences on the volume of passengers.



Field of Study:	Computer Science and Information Technology	Student's Signature .....
Academic Year:	2023	Advisor's Signature .....

## ACKNOWLEDGEMENTS

I would like to thank my advisor, Asst. Prof. Dr. Dittaya Wanvarie for her dedicated guidance, advice, and feedback throughout the process. My gratitude extends to the Airports of Thailand (AOT) for providing the data used in this research. I would like to give special thanks to the examination committees, Assoc. Prof. Dr. Nagul Cooharajanone and Dr. Ukrit Watchareeruetai, for their insightful comments and suggestions.

Sutthiya Lertyongphati



# TABLE OF CONTENTS

	<b>Page</b>
.....	iii
ABSTRACT (THAI) .....	iii
.....	iv
ABSTRACT (ENGLISH).....	iv
ACKNOWLEDGEMENTS.....	v
TABLE OF CONTENTS.....	vi
LIST OF TABLES.....	viii
LIST OF FIGURES .....	x
Chapter 1 Introduction.....	1
Chapter 2 Literature review.....	3
Chapter 3 Methodology.....	7
3.1 Proposed Framework.....	7
3.2 Dataset.....	8
3.2.1 Historical Passenger Volumes.....	8
3.2.2 Google Trend.....	10
3.2.2.1 Overview.....	10
3.2.2.2 PyTrend.....	11
3.2.2.3 Potential of Google Trend in demand prediction.....	11
3.2.2.4 Query selection.....	18
3.2.3 Meteorological characteristics.....	19
3.2.4 Events.....	23
3.2.4.1 Concerts.....	23
3.2.4.2 Sporting events.....	24
3.2.5 Econometric variables.....	25
3.2.5.1 Consumer price index (CPI).....	25

3.2.5.2 Jet fuel .....	26
3.3 Preprocessing .....	27
3.4 Regression method.....	32
3.5 Evaluation Metrics .....	33
Chapter 4 Result and discussion .....	34
4.1 Kernel Selection.....	34
4.2 Selected queries and lag order .....	35
4.3 Regression Results.....	37
4.3.1 Performance of selected queries.....	37
4.3.2 Location-specific queries .....	39
4.3.3 COVID-19 Pandemic .....	42
Chapter 5 Conclusion.....	44
5.1 Concluding the result.....	44
5.2 Future work.....	45
REFERENCES .....	47
VITA .....	51



## LIST OF TABLES

	<b>Page</b>
Table 1. Attributes of historical passenger volume data from AOT.....	8
Table 2. Data obtained for the search term "Chatuchak market" using PyTrend .....	11
Table 3. The initial set of search queries and the corresponding category .....	19
Table 4. Attribute, description, and the type of weather dataset.....	20
Table 5. Example of selected past concert from concertarchives.org and the capacity of the venues .....	23
Table 6. Examples of selected past sporting events from Wikipedia and other websites.....	24
Table 7. Number of concerts and sporting events from 2011-2020 .....	25
Table 8. Events data representation .....	27
Table 9. Performance comparison between SVR with Gaussian kernel and SVR with polynomial kernel .....	34
Table 10. Features selected by applying recursive feature elimination in each region .....	35
Table 11. Lag order of each query in different regions .....	37
Table 12. Evaluation of regression model using 10 selected features from RFE (Central region).....	37
Table 13. Evaluation of regression model using 10 selected features from RFE (Northern region) .....	38
Table 14. Evaluation of regression model using 10 selected features from RFE (Southern region) .....	38
Table 15. Comparing performance from the inclusion of additional features in the Central region.....	39
Table 16. Comparing performance from the inclusion of a location-specific feature in the Northern region .....	40
Table 17. Comparing performance from the inclusion of a location-specific feature in the Southern region .....	41
Table 18. Performance from the inclusion of dummy variables and COVID-19-related queries for the Central region.....	42

Table 19. Comparing performance from the inclusion of search volume on "Phuket sandbox" in the Southern region .....43



## LIST OF FIGURES

	<b>Page</b>
Figure 1. Proposed Framework.....	7
Figure 2. Total Thailand's inbound air passengers from 2011-2019.....	8
Figure 3. Monthly number of inbound passengers (2011-2019) .....	9
Figure 4. Daily number of inbound passengers (January 2012) .....	9
Figure 5. Google Trend Website.....	10
Figure 6. Monthly number of inbound passengers (Central region).....	12
Figure 7. Monthly number of inbound passengers (Southern region) .....	12
Figure 8. Monthly number of inbound passengers (Northern region) .....	13
Figure 9. Rising related topics in August 2011.....	13
Figure 10. Rising related queries in August 2011.....	14
Figure 11. Pearson correlation between flood-related queries and arrival passengers in the Central region.....	14
Figure 12. Monthly number of inbound passengers in all regions (2019-2021) .....	15
Figure 13. Related topics of "travel Thailand" .....	16
Figure 14. Related queries of "travel Thailand" .....	16
Figure 15. Search volume of query "Phuket Sandbox" .....	17
Figure 16. Monthly number of inbound passengers in Southern regions 2021 .....	17
Figure 17. Pearson coefficient between COVID-19-related variables and the volume of inbound passengers.....	18
Figure 18. The average temperature in all regions from 2011 to 2019.....	20
Figure 19. The average temperature in each region from 2011 to 2019.....	21
Figure 20. Average rainfall (inches) in all regions of Thailand from 2011 to 2019...21	
Figure 21. Average rainfall (inches) in each region of Thailand from 2011 to 2019.22	
Figure 22. Thailand's consumer price index (CPI) .....	26
Figure 23. Jet fuel price .....	26
Figure 24. Number of inbound passengers in Central region (February 2019).....	28

Figure 25. Daily search volume on query "Agoda Thailand" (January 2019).....	29
Figure 26. Correlation between Central inbound passengers, weather data, CPI, jet fuel prices, and events.....	29
Figure 27. Correlation between Northern inbound passengers, weather data, CPI, and jet fuel prices.....	30
Figure 28. Correlation between Southern inbound passengers, weather data, CPI, and jet fuel prices.....	31
Figure 29. Correlation between CPI, jet fuel prices, and volume of passengers in Central region.....	36



# Chapter 1 Introduction

Demand prediction in the aviation industry has always been challenging due to dynamic circumstances. Insights obtained from the forecasts could assist the airline company in various aspects, from route planning, and aircraft maintenance to revenue management. Besides the aviation industry, diverse businesses of all sizes from the government to small retailers can benefit as well. In order to achieve applicable demand prediction, besides adopting the sophisticated tool, a relevant data source must be utilized. Uncertainty and the dynamic nature of the influencing factors intensify the challenge to adapt promptly.

Factors pose an impact on each business sector differently in terms of magnitude and criticality. For instance, all businesses were disrupted by the COVID-19 pandemic as it introduces the 'new normal' which changes people's lives in all aspects. However, the disruption is directly advantageous to certain industries while some need to adapt accordingly to remain competitive. Despite most retail sales across multiple categories experiencing declines during the pandemic outbreak, some were on the opposite. With the lockdown policy limiting normal physical activities, at-home fitness equipment sales doubled while the fitness centers were negatively affected. On the other hand, some fitness centers leverage the rising use of technology in communication during the pandemic and offer online classes. Decision-making and adaptability required the ability to define relevant factors affecting individual businesses.

As previously reported in the literature, the dependency of time series analyses on a historical dataset with a consistent pattern raises doubt about forecasting accuracy when external influences are concerned (Fesenmaier et al., 2010). These influences minimize the resemblance between the past and the future. Multiple researchers explored the potential of utilizing the related factors in enhancing prediction performance, highlighting the significance of uncontrollable external factors such as weather conditions, economic variables, or the latest social trend.

As tourism and the aviation industry share a direct relationship as air travel contributes to expanding the global tourism industry, prior research on both tourism demand prediction and air passenger demand prediction utilized the volume of air passengers as one of the variables in corporations with other explanatory variables which were proved to enhance the prediction performance (Constantino et al., 2016; Fahad et al., 2013; Ghalekhondabi et al., 2019; Morley, 1994; Xiong et al., 2022) .

Most demand predictions in the aviation industry adopted extensive internal features comprising ticket prices, the number of advanced bookings, airline past performance history, and Passenger Name Records (PNR) which were demonstrated to improve the prediction performance (An et al., 2016; Liu et al., 2017; Nada et al., 2012; Vu et al., 2018). However, such data were either not publicly accessible or available for purchase at high cost by commercial providers (Mao et al., 2015). Alternatively, researchers turned to leverage data that are publicly available.

With the growth of the internet, the search engine had become an integral part of our life as it allows multiple tasks to be performed at ease, from searching for products to purchase, movies to watch, and certainly, a place to travel to. A strong correlation between the demand for several services and search queries has been recognized and proved to enhance the prediction performance to a certain degree. As its application is not limited to a specific industry, search engines have become an approachable open source of data for researchers to utilize in all domains, from retail sales to global tourism. There are many web search engine providers, namely, Google, Bing, and Yahoo.

Among the leading search engines, with over 80% of the search market share, Google is the most popular search engine on a global scale. As the most visited website, Google is undoubtedly an adequate data source for researchers.

Google Trend is an online search tool from Google that allows the user to see how often specific keywords, subjects, and phrases have been queried over a particular period. To avoid changes in the number of queries searched by an increasing number of users, the search frequency is normalized into the interval from 0 to 100.

A substantial number of previous studies have proven that the incorporation of search volume data from Google Trend can enhance the demand prediction performance in the tourism industry (Feng et al., 2019; Fesenmaier et al., 2010; Höpken et al., 2018; Kort, 2017; Li et al., 2020; Yang et al., 2015) but a few specifically perform for the aviation industry. Long et al. (2021) proposed an effective approach to identify queries to forecast the arrival of air passengers at Singapore Changi Airport. The study confirmed the potential of using Google Trend queries in air demand prediction and suggested selecting only relevant queries for better performance.

This thesis is organized as follows. In Chapter 2, the related literature review is highlighted, and the proposed framework and methodologies used to obtain data will be discussed in Chapter 3. Chapter 4 explains the process and implementation. The result and discussion are in Chapter 5 and concluded in Chapter 6. Finally, future improvement is suggested in Chapter 7.

## Chapter 2 Literature review

To date, several studies have investigated the ability to improve demand forecasting in the aviation industry. The objective of the research is beneficial to either the customer side or the airline side. Customers seek to save on cost; thus, research was conducted to find the optimal time to purchase a ticket or predict the ticket price. On the contrary, research on demand prediction could assist airlines in making decisions regarding seat utilization or flight frequency adjustment as they aim to increase revenue and keep ahead of the competition.

The majority of published studies are on ticket price prediction and passenger demand prediction based on ticket price and airline internal variables such as the number of flights (An et al., 2016; Nada et al., 2012; Vu et al., 2018). With the intention to predict ticket demand under unpredictable circumstances, Riedel et al. (2003) developed an adaptive forecasting technique incorporating the flight schedule, seasonality, and the market situation which significantly improve the forecasting performance. The study indicates that historical data should be adapted to the new situation as different influencing factors are introduced.

Customers' preferences and the market also play an essential role in demand. Liu et al. (2017) employed the Bayesian network-based topic model to uncover personal travel preferences using the Passenger Name Record (PNR), which was found to be applicable in customer travel prediction. However, acquiring data concerning the details of passengers, ticket price, or flight schedules are challenging, especially since processing personal data requires consent and ticket pricing strategies are confidential. Mao et al. (2015) mentioned the worldwide aviation data available for purchase at high-priced by commercial providers. Many researchers turned to leverage data that are publicly available, for instance, the volume of passengers.

Tourism and the aviation industry share a direct relationship as air travel contributes to expanding the global tourism industry. Prior research on tourism demand prediction and air passenger demand prediction utilized the volume of air passengers as one of the variables in corporations with other explanatory variables (Constantino et al., 2016; Ghalekhondabi et al., 2019; Morley, 1994; Xiong et al., 2022).

An et al. (2016) proposed an ensemble forecasting technique (MAP-EE) to predict a specific airline's total route demand and market share. Economic indicators such as gross domestic product (GDP), consumer price index (CPI), and personal income are also considered in the model.

External factors play a significant role in enhancing demand prediction accuracy. As previously reported in the literature, the dependency of time series analyses on a historical dataset with a consistent pattern raises doubt about forecasting accuracy when external influences are concerned (Fesenmaier et al., 2010). These influences minimize the resemblance of the past and future. Multiple researchers explored the potential of utilizing the related factors in enhancing prediction accuracy, highlighting the significance of inconsistent external factors such as weather conditions or the latest social trend.

Travelers use search engines to find relevant information for all aspects of a trip, from transportation to restaurants to accommodation. Thus, these queries

reflect the trend in their preference and predict their future behavior (Yang et al., 2015). The concept can be extended to further improve demand prediction strategy, not only for air passengers but also in other markets such as hotels and other tourism-related businesses.

Yuan (2014) proved that by utilizing external factors, specifically travel-related queries extracted from search engines, the performance of the ticket sales prediction of an online travel agent company is enhanced. Emphasizing that one should not rely only on historical data.

Kort (2017) utilized Google Correlate, a data source provided by Google, in search for related search terms with a similar pattern to the input dataset. Unfortunately, Google Correlate was discontinued in 2019. However, a similar service, Google Trend is still active when this research is being conducted.

Multiple types of research prediction used Google Trend as one data source and proved to be applicable in numerous areas. Google Trend was utilized by Fu et al. (2022) to inspect occupant behavior and predict building energy consumption and Amusa et al. (2022) in modeling the recent COVID-19 incidence.

The incorporation of data from Google Trend and machine learning techniques was proven to enhance the demand prediction performance substantially in the tourism industry (Feng et al., 2019; Fesenmaier et al., 2010; Höpken et al., 2018; Kort, 2017; Li et al., 2020; Yang et al., 2015). Specifically, in the aviation industry, Long et al. (2021) introduced the application of the Neural Granger causality model to select the relevant search terms from Google Trends for air passenger forecasting at Singapore Shangi Airport. The study had proven that using only selected search term outperformed the inclusion of all terms acquired from Google Trend. However, the author did not cover the time period after COVID-19 pandemic.

Destination Insights, a recent service provided by Google, was explored by Rashad (2022) aiming to improve the forecasting model of the tourism demand. Destination Insights allows the user to view data related to travel demand of the designated origin and destination within a specific date range. The result of the research confirmed that the Google-Augmented Model performs better than the traditional model. Despite the feasibility, the data is only available from 2020 onwards, thus, it is not applicable if the historical data prior to the service is taken into account.

Raising the impact analysis of events on demand prediction, previous studies have almost exclusively focused on those that introduced the negative effect on demand, namely natural disasters, pandemics, and terrorists (Prideaux et al., 2006). This research seeks to look from another perspective and study the impact of events in terms of sporting events and music festivals. Contrasting the unforeseen circumstance of natural disasters, events are planned in advance. However, they do not occur seasonally and may assist in short-term demand prediction. As reported by Adjallah (2022), the recent music festival “Tomorrowland” draws a large number of visitors from over 60 countries and over 1.3 million visitors are expected to arrive in Qatar to attend World Cup 2022 according to Olmsted (2022). This indicates an increasing trend in people traveling abroad to attend events. Although the attractiveness of the events varies,



there is a potential to consider them as one of the external variables in forecasting air passenger demand.

Another influencing variable that plays a significant role in aviation is the weather. It is mostly concerned with the sense of safety and continuity of flights. Multiple researchers provided evidence of their impact on flight delay and cancellation prediction and particularly, in the tourism industry (Choi et al., 2016; Scott et al., 2010). Besides the fact that extreme weather conditions evidently affect the air passenger demand, usual conditions like rain can also impede the decision to travel. Mao et al. (2015) discovered that the temperature at the destination affects the number of passengers at a particular time of the year. Thus, the impact of the weather on Thailand's inbound passengers will be justified in this research.

With the major contribution of the tourism and aviation industry to the economic growth rate, extensive literature has developed models with an econometric approach in demand forecasting, and the result has been shown to outperform the univariate conventional model (Ghalekhondabi et al., 2019; Vu et al., 2018). The macroeconomic factors broadly included are the consumer price index (CPI), gross domestic product (GDP), and exchange rate. The relationship between CPI and tourism had been confirmed by Morley (1994) as employing CPI data could improve demand prediction. Combining CPI and exchange rate, Kim et al. (2017) indicated that the prices of tourism products are the most important factor in the choice of destination by tourists.

Aside from CPI, Wang et al. (2019) introduced crude oil prices as one of the explanatory factors in airfare price prediction and conclude that distance, seat class, and passenger volume are the most prominent features, but to achieve higher prediction accuracy, other less important factors such as crude oil must be included. However, Kim et al. (2017) highlighted the fact that crude oil prices may have a different trend from jet fuel used by aircraft. On average, 30% of an airline's operating costs are from the fuel cost, thus, an increase in fuel cost could affect the airfare and consecutively influence the passenger's decision to fly. Jet fuel data is publicly available online and will be included as an explanatory variable in this research.

As traveling by air dominates international tourism with an increasing trend (World Tourism Organization, 2021). The demand for aviation and tourism may share similar influencing factors. The literature in search of major influences on air travel demand and tourism demand prediction draws contrasting conclusions according to differences in the country of subject matter, in terms of both countries of origin and destination. National income and labor force are concluded to be the major influences on Kuwait's air travel demand (Fahad et al., 2013) while CPI and exchange rate are more fundamental in Mozambique's tourism demand (Constantino et al., 2016). Concluding that the particular influences may be only legitimate to the country of research. To our knowledge, limited research has been conducted to predict air passenger demand in Thailand.

Moreover, the selection of features included in the experiment differs. As it had been proven that utilizing more than one feature can achieve higher prediction performance (Constantino et al., 2016; Varian et al., 2009; Vu et al., 2018). Previous studies on demand prediction have almost exclusively focused on

a certain group of contributing factors. A limited amount of research has been conducted that simultaneously considers Google Trends, weather conditions, econometric variables, or other related influences. With historical data on air passenger volume provided by Airports of Thailand (AOT), this paper will explore the impact of several influencing factors on inbound air passenger demand in Thailand.



## Chapter 3 Methodology

### 3.1 Proposed Framework

The framework of this research is illustrated in Figure 1. The historical volume of air passengers from Airports of Thailand (AOT) collected on a daily basis is used for modeling.

The process starts with data collection of the explanatory variables. An initial set of keywords will be selected, and the daily search volume of each keyword is congregated from Google Trends.

After Thailand's historical weather data, consumer price index (CPI), and jet fuel price are downloaded, all datasets are preprocessed. Gradient Boosting Regression, Random Forest Regression, and Support Vector Regression are utilized to predict passenger demand with and without the inclusion of influencing factors obtained, concluding which factor has the most impact.

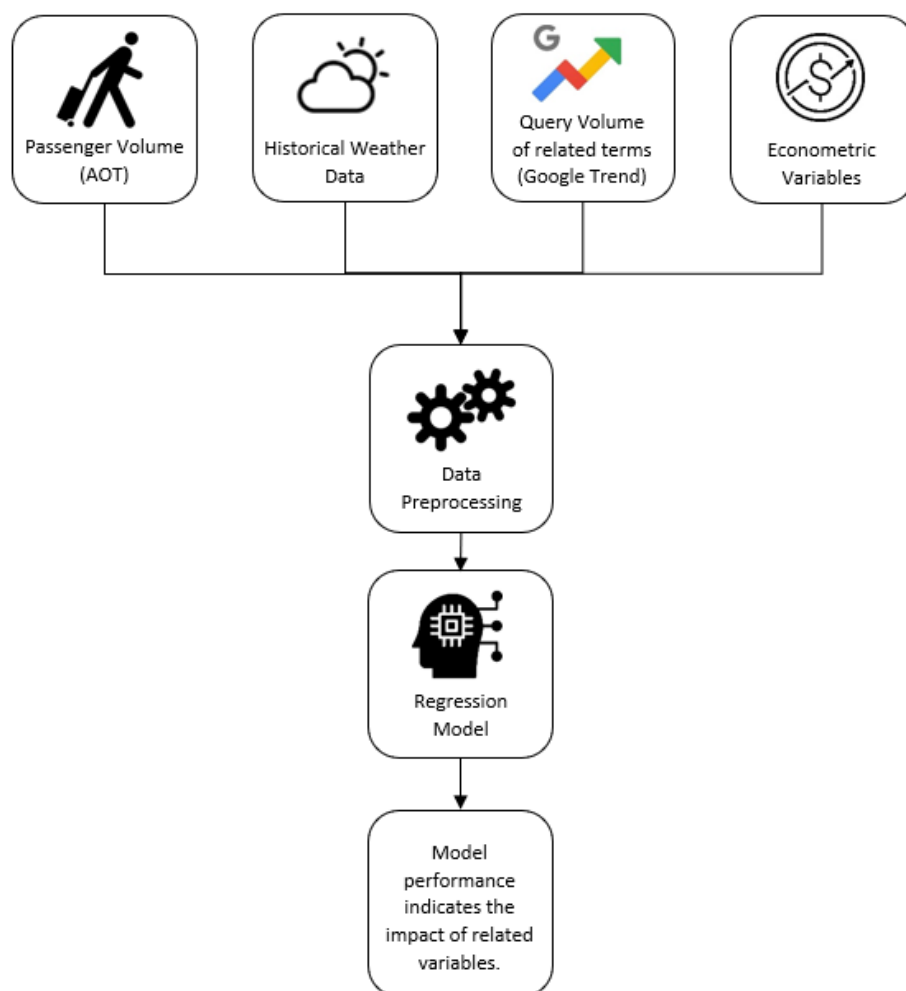


Figure 1. Proposed Framework

## 3.2 Dataset

### 3.2.1 Historical Passenger Volumes

The number of passengers which is the target output in this research is provided by Airports of Thailand (AOT). AOT manages Thailand's six international airports located in different regions. Bangkok (BKK) and Don Muang (DMK) are in the Central region. Chiang Mai (CNX) and Chiang Rai (CEI) are in the Northern region. Phuket (HKT) and Hat Yai (HDY) are in the Southern region.

The data consists of daily aggregated tourist arrivals and departures, collected from 2011 to 2021. This research will focus on the inbound passengers; thus, the originating country will be disregarded. The six airports will be consolidated according to their corresponding regions as they share similar weather conditions. The features used in this research are shown in Table 1.

Attribute	Definition	Type
AIRPORT_CODE	Arrival Airport Code	Object
ACTUAL_DATE	Arrival Date	Object
PAX_TOTAL	Number of arrival passengers	Integer

Table 1. Attributes of historical passenger volume data from AOT

As illustrated in Figure 2, briefly observing the overall trend of Thailand's inbound passengers from 2011-2019 in six operating airports, excluding the data after the emergence of COVID-19 pandemic in 2020, the highest number of inbound passengers is in January, and the lowest is in June. There was a slight decrease the September before the number started to rise again until January.

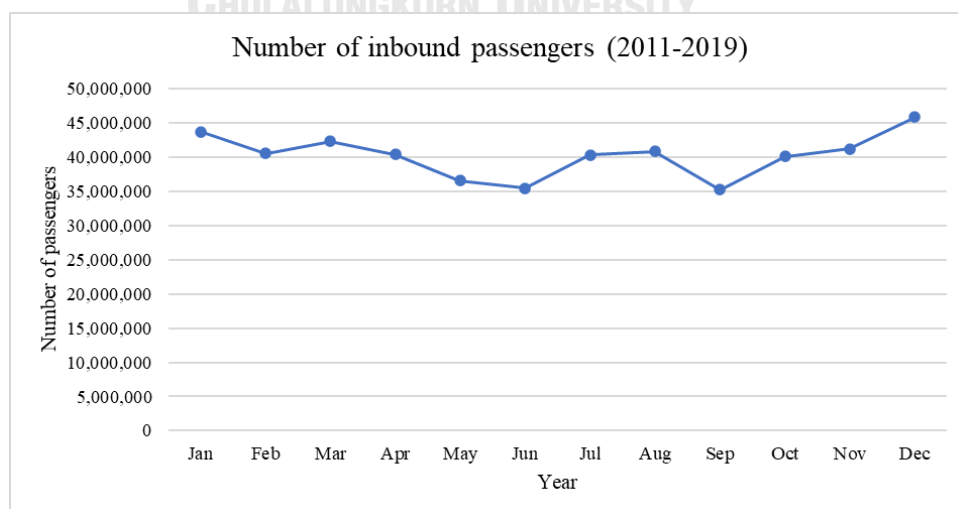


Figure 2. Total Thailand's inbound air passengers from 2011-2019

Regardless of the number of passengers, the monthly trend from 2011-2019, plotted in Figure 3, all regions share a similar rise and fall. However, daily trends, for instance, in January 2015 as plotted in Figure 4, have shown more unique fluctuations. Thus, the study will investigate each region individually.

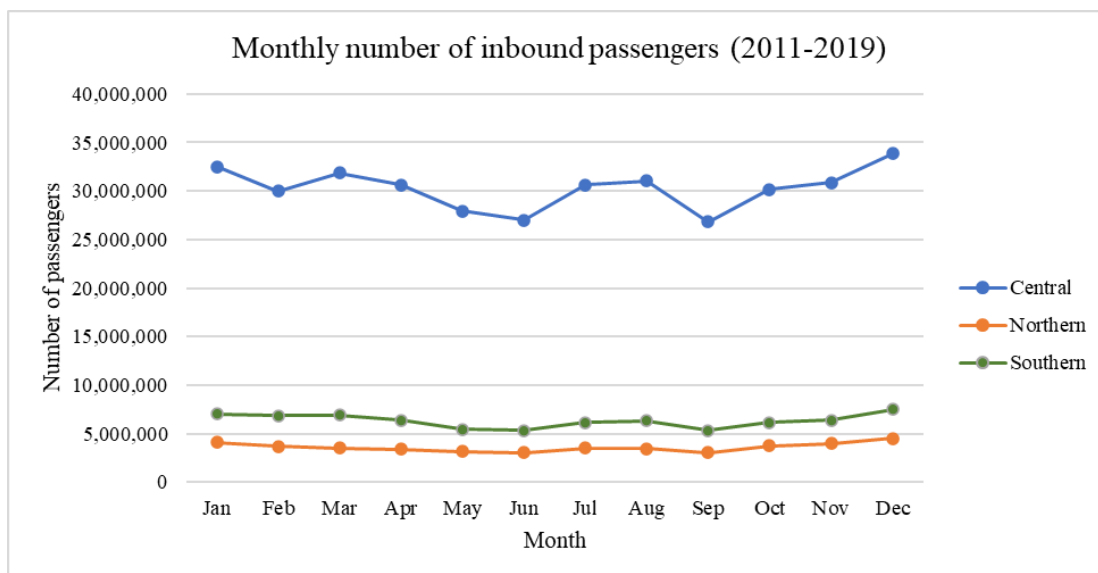


Figure 3. Monthly number of inbound passengers (2011-2019)

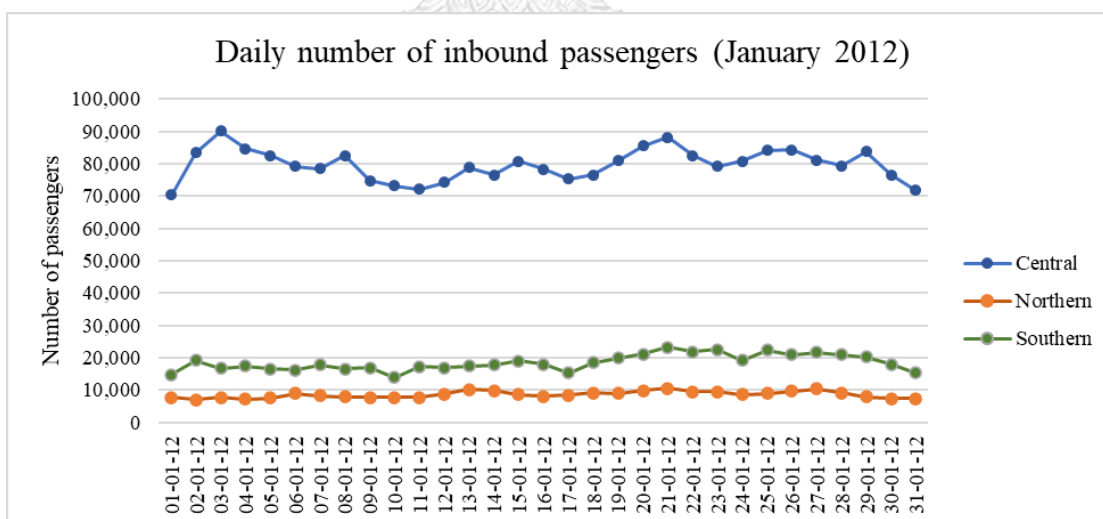


Figure 4. Daily number of inbound passengers (January 2012)

## 3.2.2 Google Trend

### 3.2.2.1 Overview

Google Trend is an online search tool that allows the user to see how often specific keywords, subjects, and phrases have been queried over a particular period. To avoid changes in the number of queries searched by an increasing number of users, the search frequency is normalized into the interval from 0 to 100.

The user can type in the query via <http://www.google.com/insights/search> to generate a graph illustrating the interest over time. The numbers representing search interest relative files can be directly downloaded in CSV format. Figure 5 below illustrates the example of interest over time for the search term “travel Thailand”. It should be noted that Google is not case sensitive according to the fact that users often search in lowercase and the search result should illustrate the same context regardless of the case. Misspellings, plural, or singular versions of the search terms are not included in the result.

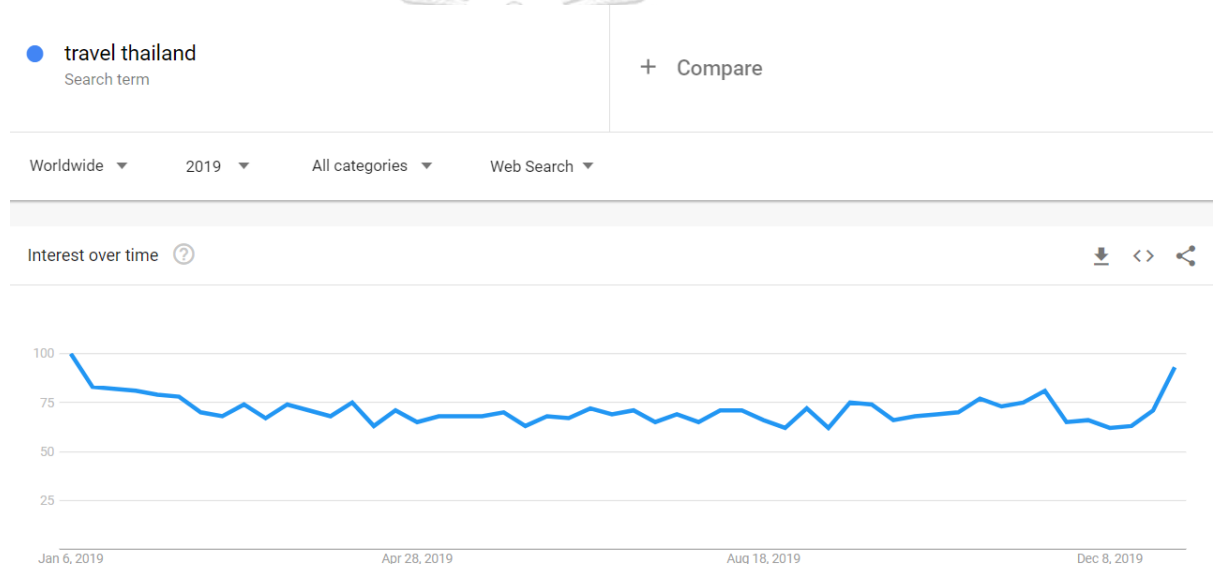


Figure 5. Google Trend Website

The search result can be explored according to the country of the searchers or select “Worldwide” to capture global interest. The time period can be specified in different window sizes, namely, years, months, days, and hours, with the latter only available for data in the past week.

The data can also be filtered according to the categories, for instance, “Business & Industrial” or “Law & Government”. The user can specify if the search is made through a normal “Web Search”, “Image Search”, “News Search”, “Google Shopping” or “YouTube Search”.

It should be highlighted that the search data is indexed according to the specified window size. Contrasting results can be obtained if the same word is searched in a different time frame. Moreover, the frequency of the data point depends on the specified window size.

For an equivalent search term, the search results in the year range will be provided in monthly frequency as opposed to results for a single year which will be on a weekly basis. Additionally, daily data points can be obtained by stating a single month. Thus, the appropriate window size should be selected according to the user's requirements.

### 3.2.2.2 PyTrend

Searching each keyword individually can be time-consuming. Alternatively, to extract the data faster and more conveniently, an open-source Python package called PyTrend is used. PyTrend is a free pseudo-API for Google Trends which allows the interface to automatically download reports from Google Trends. The algorithm has been implemented with PyCharm (a Python Integrated Development Environment). The parameters defined are equivalent to those available in the web interface.

Date	Chatuchak market_unscaled	Chatuchak market_monthly	isPartial	scale	Chatuchak market
01-01-2011	49	85	FALSE	0.85	41.65
02-01-2011	32	85	FALSE	0.85	27.2
[...]	[...]	[...]	[...]	[...]	[...]
01-02-2011	29	82	FALSE	0.82	23.78
02-02-2011	34	82	FALSE	0.82	27.88
[...]	[...]	[...]	[...]	[...]	[...]

Table 2. Data obtained for the search term "Chatuchak market" using PyTrend

Table 2 shows the data obtained for the search volume of the term "Chatuchak" using PyTrend. The second column is the data fetched month by month and is comparable within a month but not throughout the intra-month. The following column contains data in monthly frequency as they are obtained in the year range. The "scale" column is the monthly search interest weight for each month. For the unscaled data to be comparable through time, it is multiplied by the scale and the result is acquired in the last column.

The terms searched by users who searched the given input term can be viewed as related topics and related queries. A "topic" is a group of terms that share the same concept in any language, while a "Search term" only includes data for that language. They can be sorted by the quantity and time of the search. Top searches are the most searched queries in a specific time frame. Rising searches are those that are accelerating fast.

### 3.2.2.3 Potential of Google Trend in demand prediction

However, to effectively selected the related topic and queries, the period should be within a month since top searches and queries may vary between

months throughout the year. For instance, related topics and queries on “flooding” are obtained after July due to severe flooding that occurred during the monsoon season.

The volume of inbound passengers in each region from 2011 to 2019 is plotted in Figure 6,7 and 8. 2020 and 2021 were excluded from the comparison as the trends abruptly change due to the COVID-19 pandemic. The latter case will be investigated separately.

As shown in Figure 6,7 and 8, the number of arrival passengers in all regions is in a decreasing trend from August to September due to the monsoon season and started to climb towards the end of the year. However, a unique decline can be observed between October and November 2011. The falls was minimal in the Southern and Northern region compared to the Central region.

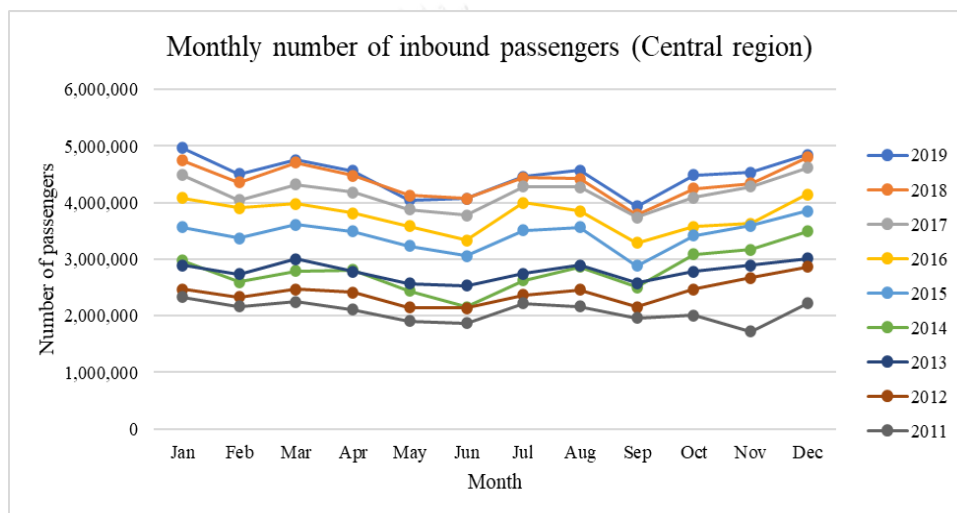


Figure 6. Monthly number of inbound passengers (Central region)

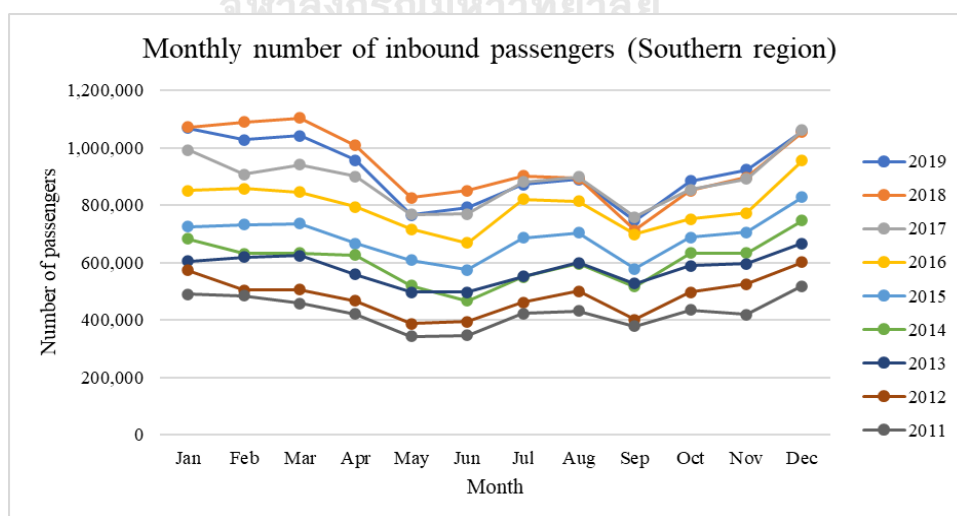


Figure 7. Monthly number of inbound passengers (Southern region)



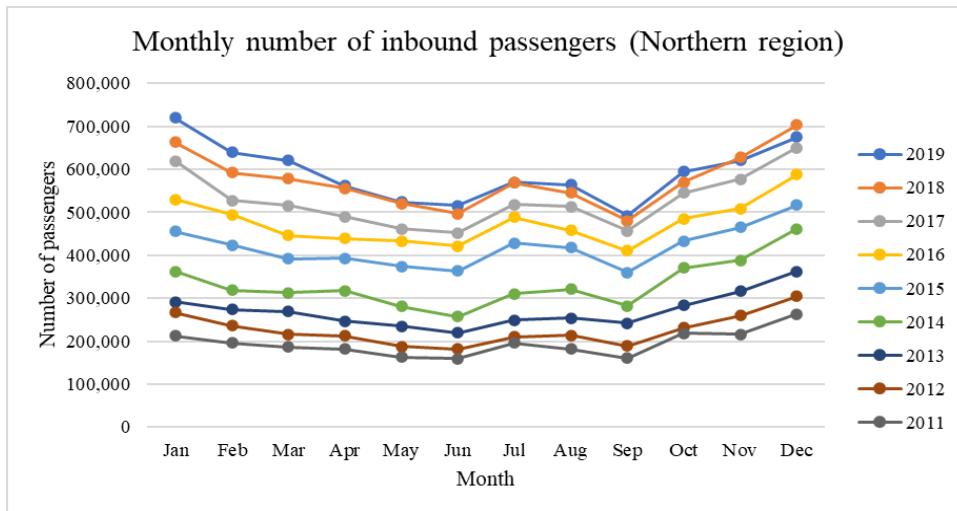


Figure 8. Monthly number of inbound passengers (Northern region)

Without prior knowledge of the reason behind the decline, Google Trend can assist in clarification. With the initial search keyword “travel Thailand”, related topics and queries on “flooding” started to rise in August as shown in Figures 9 and 10. The “Breakout” shown next to the queries indicates that the search term rose higher than 5000% compared to the previous term period.

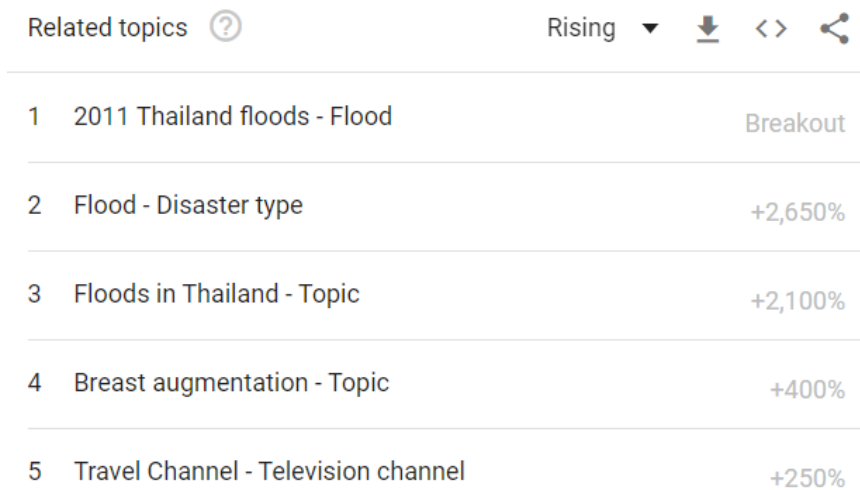


Figure 9. Rising related topics in August 2011

Related queries ? Rising ▼ ↓ <> ↻

---

1	bangkok floods	Breakout
2	flooding in thailand	Breakout
3	travel expo thailand 2011	Breakout
4	thailand floods 2011	Breakout
5	travelodge	Breakout

Figure 10. Rising related queries in August 2011

The flood started in the Northern region before severely affecting Bangkok in October. Figure 11 indicates the negative correlation between flood-related queries and arrival passengers in the Central region. Ability to capture and monitor such queries and determine the possible impact on the trend can assist in short-term demand forecasts.

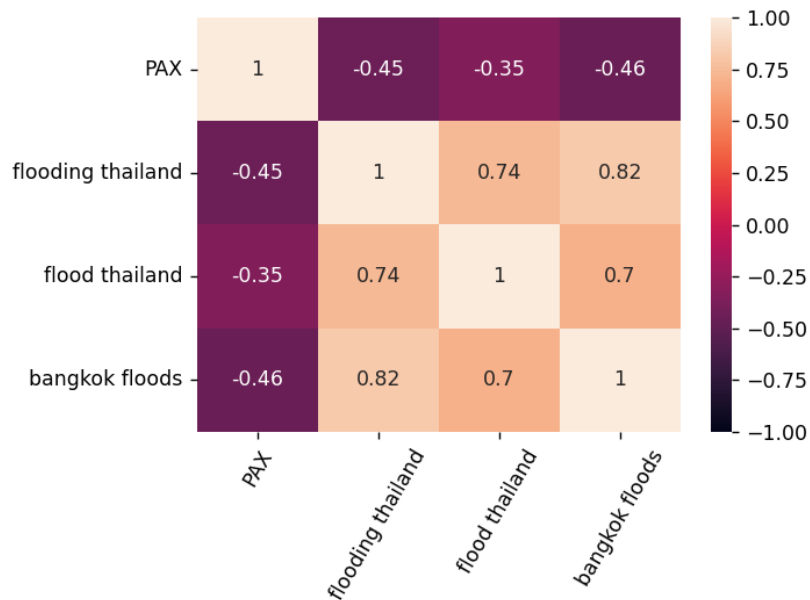


Figure 11. Pearson correlation between flood-related queries and arrival passengers in the Central region

Aside from the assumption that concerts may introduce a sudden increase in the volume of passengers, a drastic decline is clearly observed in Figure 12 after 2019 explicitly due to the COVID-19 outbreak. The pandemic resulted in travel restrictions across the globe.

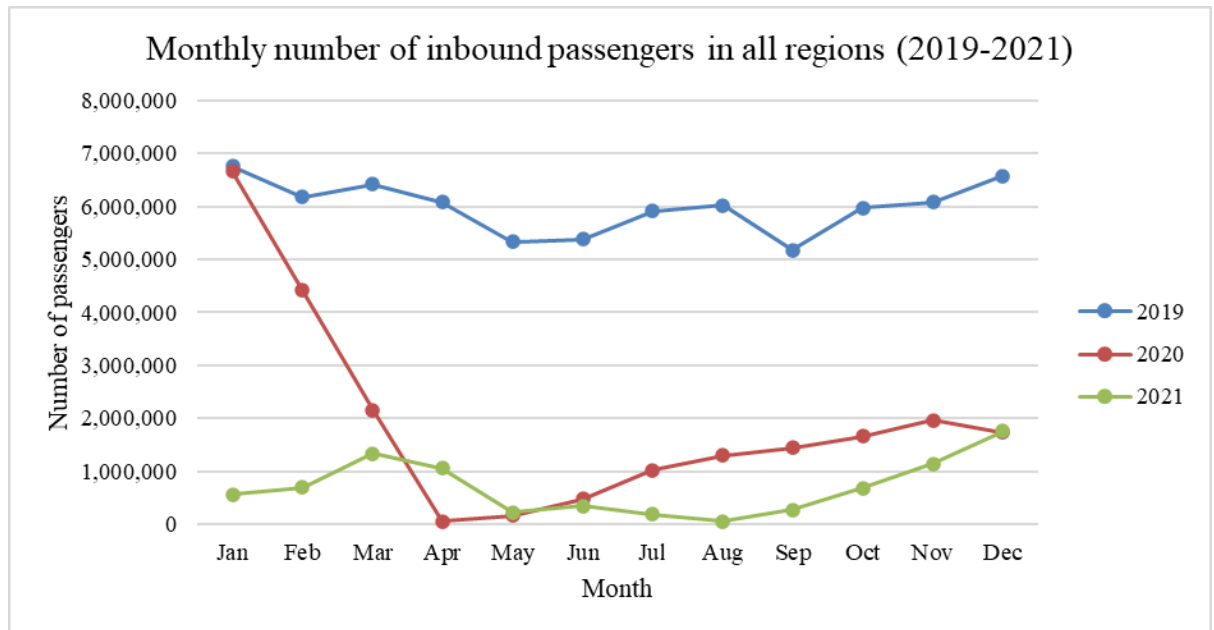


Figure 12. Monthly number of inbound passengers in all regions (2019-2021)

Kim et al. (2017) used dummy variables to represent the occurrence of special events such as the Fukushima nuclear disaster in tourism demand prediction. The dummy variables were coded as zero before the occurrence and one afterward. Thus, in this research, a dummy variable for COVID-19 coded as one in will be included from 2020 onwards in order to estimate the impact of the pandemic on the volume of inbound passengers.

According to Wikipedia, the first case of COVID-19 outside China was first detected in Thailand on 13 January 2020. Similar to the flood incident in 2011 when related queries on flood started to rise after the occurrence, people who searched for “Travel Thailand” started to search for COVID-19-related queries as well at the beginning of January 2020. The rising related topics and queries as shown in Figures 13 and 14 remained similar throughout the year. Until mid-2021 when queries on “Phuket Sandbox” and “Thailand travel restriction” started to rise.

Related topics 		Rising    
1	Coronavirus - Virus	Breakout
2	Coronavirus disease 2019 - Disease	Breakout
3	Virus - Infectious agent	Breakout
4	Severe acute respiratory syndrome coronaviru...	Breakout
5	Foreign, Commonwealth and Development Offi...	Breakout

Figure 13. Related topics of "travel Thailand"






Related queries 		Rising    
1	coronavirus thailand travel	Breakout
2	coronavirus	Breakout
3	travel to thailand coronavirus	Breakout
4	is it safe to travel to thailand coronavirus	Breakout
5	coronavirus in thailand	Breakout

Figure 14. Related queries of "travel Thailand"

“Phuket Sandbox” refers to the model officially launched in late June, allowing foreign travelers to enter Thailand without quarantine under the condition that they must stay within Phuket, an island located in the South, no less than 5 nights and get a negative COVID-19 test result before traveling elsewhere.

The search volume for the query “Phuket sandbox” is illustrated in Figure 15. The search volume is high in June as people may have heard the announcement and started looking for a possibility to travel to Thailand. However, Figure 16 indicates an increasing number of passengers arriving in the Southern after August. As travelers tend to search for information months before the actual departure, such queries have the potential in assisting the prediction of short-term future demand.

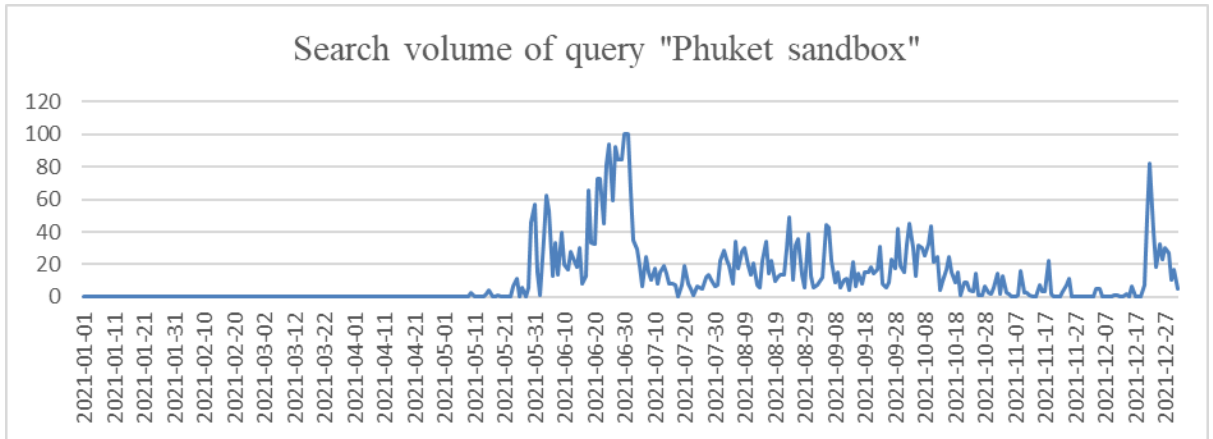


Figure 15. Search volume of query "Phuket Sandbox"

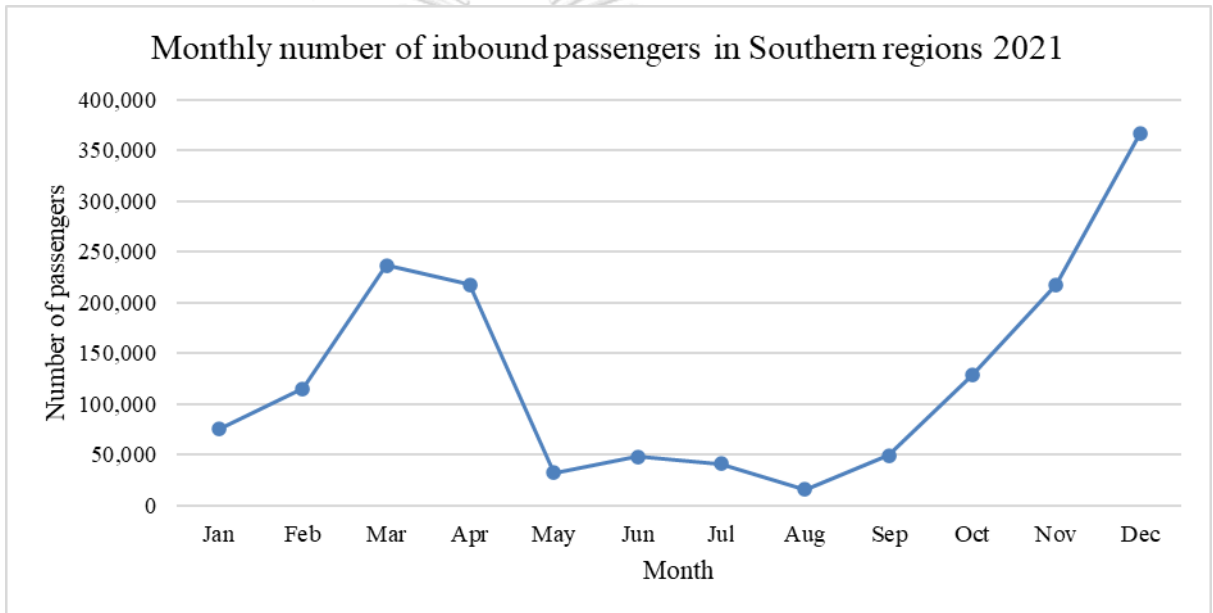


Figure 16. Monthly number of inbound passengers in Southern regions 2021

As shown in Figure 17, Pearson correlation established a negative correlation between COVID-19 dummy variables, the search volume of the related queries “coronavirus” from Google Trend, and the volume of passengers. The dummy variable showed a higher negative correlation than the search volume. However, the inclusion of other queries such as the aforementioned “Phuket sandbox” could further enhance the prediction result as after the steepest drop in January 2019, the volume of passengers fluctuates with the changes in travel restrictions in the originating countries and Thailand. Travelers who performed the search may discover updates for the travel restriction and decide to take a trip.

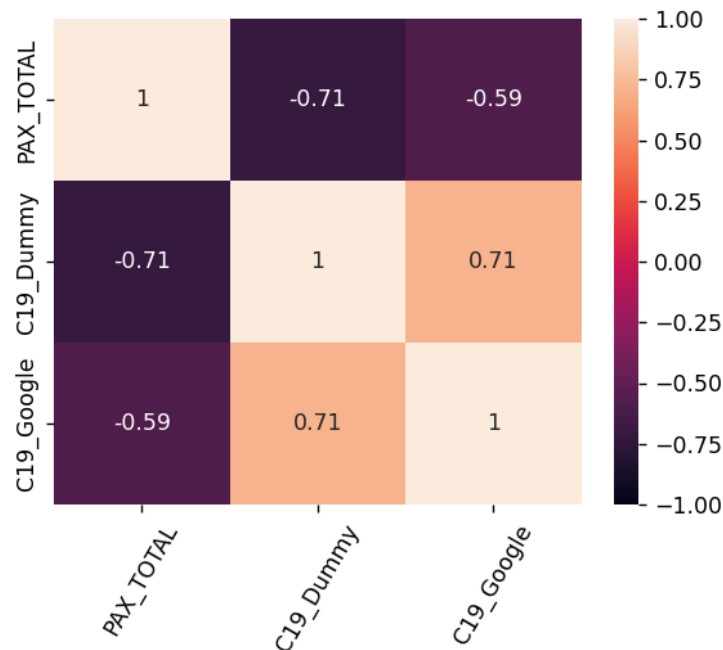


Figure 17. Pearson coefficient between COVID-19-related variables and the volume of inbound passengers

#### 3.2.2.4 Query selection

In accordance with the passenger data obtained, the related query must be selected upon the type of passenger and the country of destination, which in this study is Thailand. According to the statistic (National Statistical Office of Thailand, 2022), the majority of the inbound passenger to Thailand is leisure travelers. There is an increasing trend over the years and accounts for over 90% in 2019 compared to other categories and remained predominance although the trend declined from the COVID-19 pandemic. Thus, the queries are preliminarily selected for the major type of travelers. Additionally, from statistics, some business travelers planned to extend a trip for leisure purposes as well, the selected query may also reflect their demand.

The query used when searching about traveling is often specific to the destination. Initially, 12 keywords are selected to reflect various aspects of the trip including transportation, location, and accommodation-related queries as shown in Table 3. The terms are chosen according to previous studies (Höpken et al., 2018; Li et al., 2020) and the author's knowledge and judgment. Google Trend provides related topics and related queries which are also searched by the user searching the initial term. Topics in Google Trends are a set of terms that share an identical concept across diverse languages, while queries uniquely match the language of the user's input (Höpken et al., 2018). Since this research does not focus on any specific originating country, the language used by the user when searching is diverse. To reduce the language boundary, both related topics and

queries are explored and selected according to the initial search term. The initial search terms are listed in Table 3 with the corresponding categories.

Category	Search Query
Accommodation	Agoda Thailand
	Hotel Thailand
	TripAdvisor Thailand
Currency	Thai Baht
Places	Shopping mall Thailand
	Thailand temple
	Places visit Thailand
Transportation	Map Thailand
	Visit Thailand
	Airfare Thailand
	Travel Thailand
	Flight Thailand

Table 3. The initial set of search queries and the corresponding category

### 3.2.3 Meteorological characteristics

To prove whether local weather condition impacts passenger traveling by air or not, Thailand's historical weather data is necessary. Thailand's historical weather data is publicly available for download on the website of the Thai Meteorological Department or specific weather variables of choice within date ranges can be requested by email.

However, the data obtained contains multiple missing values in particular months. In awareness of inaccurate predictions from using incomplete data, historical weather data is purchased from <https://www.worldweatheronline.com/>. The website provides weather forecasts for destinations around the world with more refined weather elements. The dataset contains no missing values. Thus, additional missing data handling is not necessary. Table 4 below illustrates the attributes.

Attribute	Description	Type
year	Year	Integer
month	Month Index	Integer: 1-12
avgtempC	Average temperature in degrees Celsius	Float
Rain Inches	Average monthly rainfall in inches	Float
rain_days	Number of days with rainfall	Integer
avgwindspeedKmph	Average Wind Speed in Kmph	Float
visibilityKm	Visibility in km	Integer
humidity	Average humidity (%)	Integer
uvindex	Average UV Index	Integer
sun_hour	Average Sun Hour	Float
sun_days	Average Sunny Days	Float

Table 4. Attribute, description, and the type of weather dataset

The temperature is in an increasing trend from February to May and reaches its peak in April before declining when the rainy season arrived in May. In Figure 18, the temperature in all regions shares a similar trend in contrast to the amount of rainfall that differs regarding the regions as shown in Figure 19.

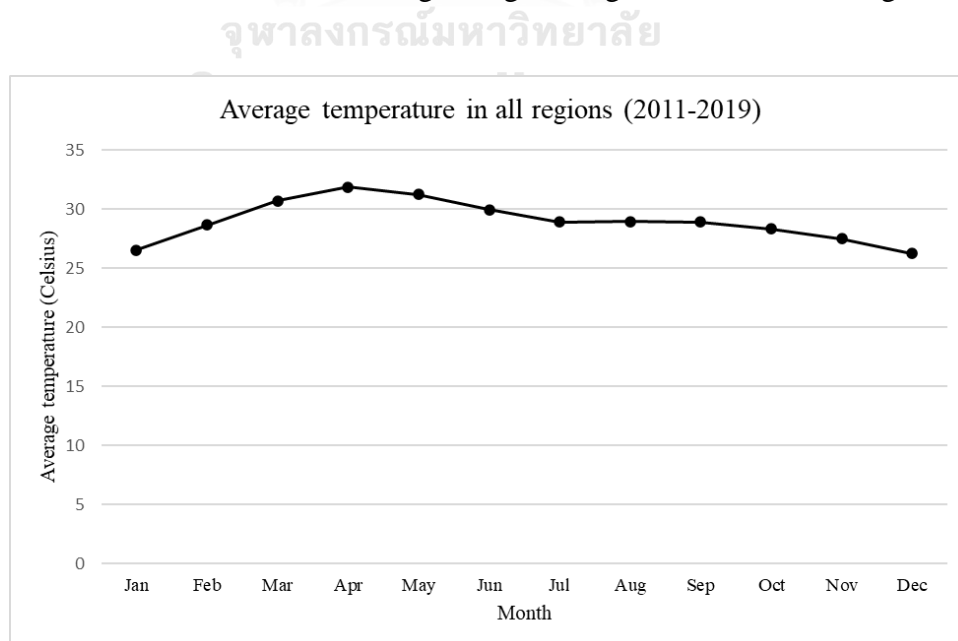


Figure 18. The average temperature in all regions from 2011 to 2019



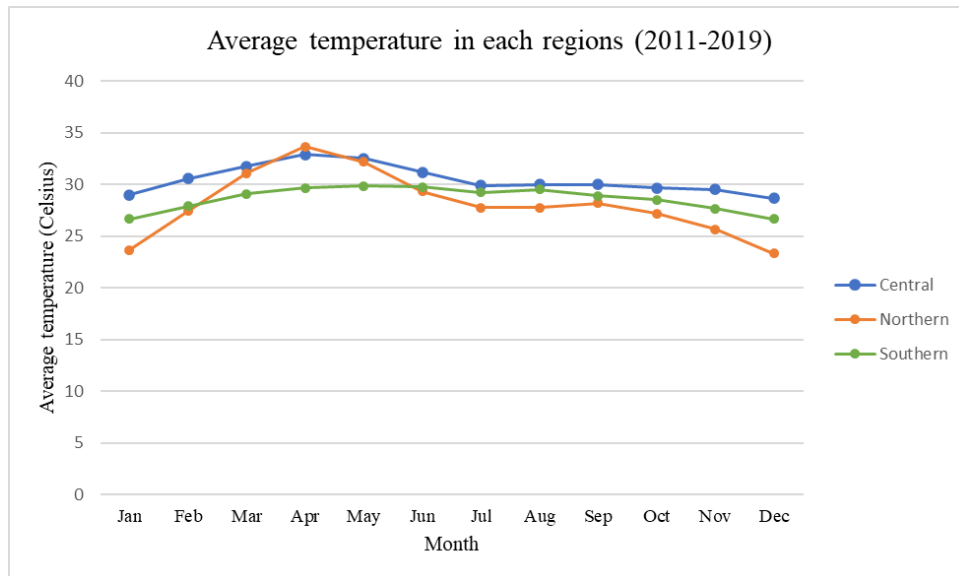


Figure 19. The average temperature in each region from 2011 to 2019

Figures 20 and 21 illustrate the average rainfall in all regions of Thailand and in each region from 2011-2021 respectively. Average rainfall in all regions is lowest in February and rises until May before declining and increasing to reach its peak in October. However, the trend is different when observing each region separately. For the Central region, the highest rainfall is in October, while the peak value for Northern and Southern is in August and November respectively.

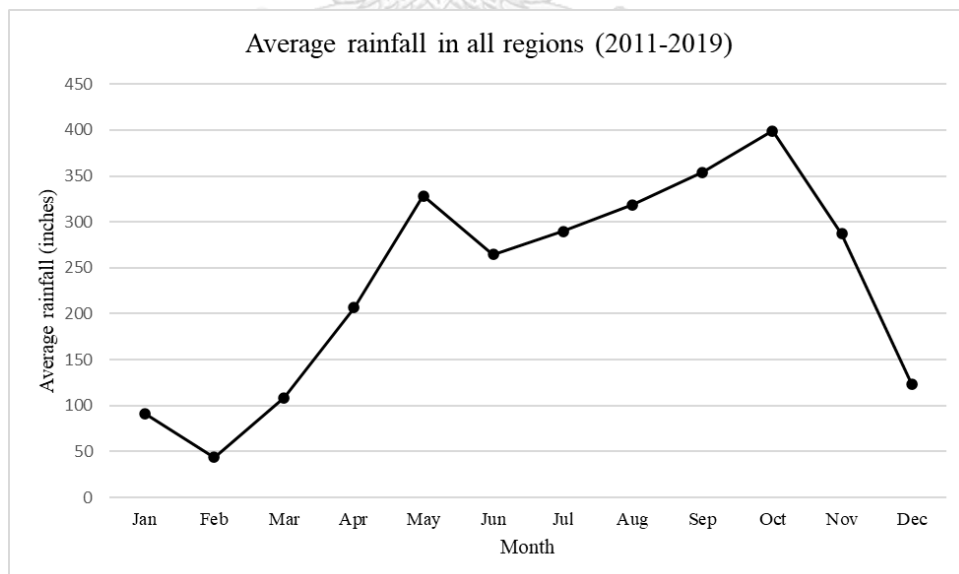


Figure 20. Average rainfall (inches) in all regions of Thailand from 2011 to 2019

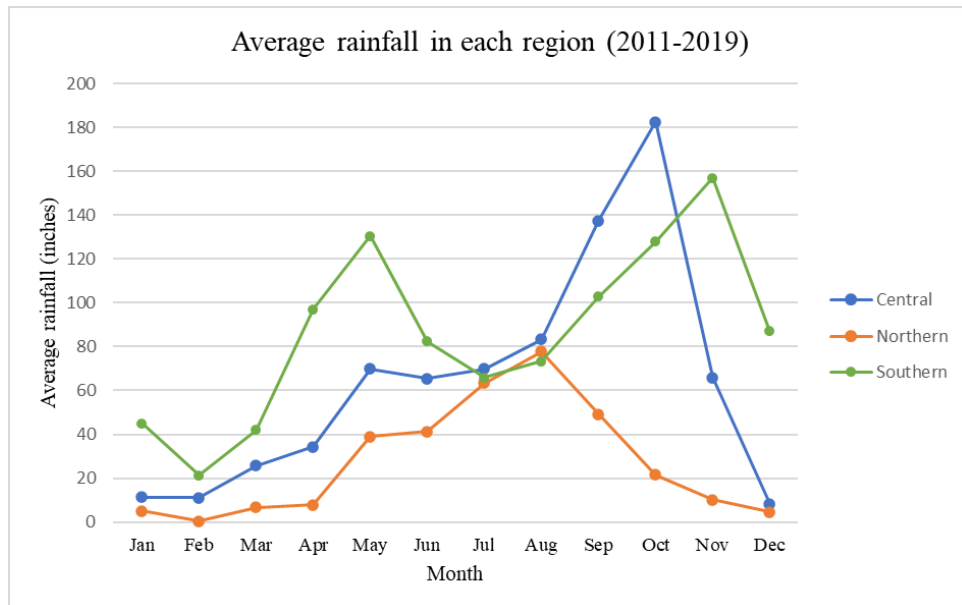


Figure 21. Average rainfall (inches) in each region of Thailand from 2011 to 2019

As the weather condition is not aligned in different regions of the country, they will be investigated individually. Travelers who intended to travel to each region may be expecting different weather conditions. For instance, high humidity could be preferable if the plan is to visit mountains in the Northern as they may expect to witness the sea of mist that occur relative to the humidity level.

### 3.2.4 Events

Events considered in this research are concerts and sporting events. The following section will discuss the data collection of both categories.

#### 3.2.4.1 Concerts

The historical record of the concerts was manually collected from <https://www.concertarchives.org/> which provide a list of past concerts in Thailand with detail on dates and venues.

The selected events are based on location and the size of the venues. Only international artists are chosen in particular as local artists are less likely to attract fans at the global level. There is no record on the number of attendees for every concert held in Thailand, therefore, the list of highest grossing concert tours provided by Wikipedia is used as a reference in the selection process. According to the website, dated back to 1977, most concerts are held in Bangkok with a total of over 400 concerts and less than 10 concerts in other regions. Thus, this research specifically considered concerts in Bangkok.

The dates of the concerts range from 2011 to 2019 as most concerts are canceled or postponed due to the COVID-19 pandemic. Table 5 shows an example of selected concerts with the size of the venue.

Date of Concert	Name of the Concert/Band/Artist	Venue	Capacity
2011-04-23	Maroon 5	Impact Arena	12,000
2011-03-10	Slash	Bangkok Convention Centre (BCC)	2,000
2011-09-23	Linkin Park	Aktive Square	20,000
2012-05-25	Lady Gaga	Rajamangala National Stadium	49,749
2015-10-01	The Chainsmokers	Onyx	2,000
2016-06-11	Road to Ultra Thailand 2016	Bitech Bangna	7,000
2017-04-13	S2O 2017	SHOW DC OASIS ARENA	25,000
2017-08-17	All-4-One	Thunderdome Complex	20,000

Table 5. Example of selected past concert from concertarchives.org and the capacity of the venues

With insufficient data on the number of attendees and the artist's popularity being difficult to measure, the concerts held in a venue that could accommodate more than 2,000 people were selected in this research. Concerts with a low number of attendees may not introduce a significant impact on the number of inbound air passengers.

### 3.2.4.2 Sporting events

Sporting events are also manually collected from Wikipedia. As the capacity of the venue is not specified, the sporting events are selected based on the size of the competition, specifically at an international level in the assumption to attract oversea attendees more than the domestic competition.

For instance, 600 athletes from 87 nations attended in 2019 IFMA World Muaythai Championships held at Huamark Sports Complex. The number of attendees is assumed to be higher as apart from participants, the presence of spectators or coaches is expected as well.

Some events, such as marathons, were not listed as international sports competitions in Wikipedia are discovered manually by searching for past sporting events in Bangkok. “Bangkok Marathon”, the largest traditional marathon event in Thailand held annually in Bangkok since 1987, was found to attract international runners globally. Although the number of participants was not specified each year, approximately 3,000 international runners from 60 countries were found to attend Bangkok Marathon 2019. Thus, Bangkok Marathon is included as one of the sporting events in November from 2012 to 2019, none in 2011 and 2016 but twice in 2012 and 2017 respectively due to the postponement. An example of selected sporting events is shown in Table 6.

Date	Name of Competition
2012-11-01	2012 FIFA Futsal World Cup
2012-12-14	2012 Race of Champions
2013-03-16	2013 IIHF Challenge Cup of Asia
2017-04-01	2017 Youth World Weightlifting Championships
2019-07-22	2019 IFMA World Muaythai Championships
2019-11-17	Bangkok Marathon 2019

Table 6. Examples of selected past sporting events from Wikipedia and other websites

Based on the condition, a total of 84 concerts and 21 sporting events held in Bangkok were selected. Table 7 shows the number of selected concerts, and sporting events in each month from 2011 to 2020.

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Monthly
<b>Jan</b>			1		1			1		1	4
<b>Feb</b>	3	2				2	2	1	1		11
<b>Mar</b>	1	1	1	1	2	2	1	1			10
<b>Apr</b>	2			1			3	3	3		12
<b>May</b>		1	1		1		1	2	3		9
<b>Jun</b>	1	1				1	2				5
<b>Jul</b>			1			2		1	3		7
<b>Aug</b>	1	1			2	1	2		2		9
<b>Sep</b>	1	1	1		2	1	1	1	3	1	12
<b>Oct</b>		1		1	1			1	2	1	7
<b>Nov</b>		2	2	1	3		2	1	3		14
<b>Dec</b>		1				1	2	1			5
<b>Yearly</b>	9	11	7	4	12	10	16	13	20	3	105

Table 7. Number of concerts and sporting events from 2011-2020

### 3.2.5 Econometric variables

Thailand has a mix-typed economy. The majority is based on industry, tourism, service, and natural resources (National Statistic Office Ministry of Digital Economy and Society, 2021). Specifically, the tourism industry draws large income for Thailand with numerous tourist attractions and services offered.

In accordance with the economic demand theory indicating demand for goods and services and their corresponding prices share a negative correlation, the attractiveness of the destination may increase as the travel cost is lower (Eglitis, 2020). Thus, previous research often includes tourism prices as one contribution to demand prediction (Uzama, 2009)

Not only the direct cost such as the transportation cost for traveling to the destination but also the cost of travel during the stay. Due to the absence of data on tourism prices, consumer index price (CPI) is alternatively used in demand prediction (Kim et al., 2017; Morley, 1994).

#### 3.2.5.1 Consumer price index (CPI)

The consumer price index (CPI) is the price of a weighted average market basket of consumer goods and services purchased by households. CPI measures the change in the price of goods and services from the consumer's perspective. CPI includes prices across multiple categories namely food, clothing, shelter, and fuels; transportation fares; service fees. Pektaş (2020) indicates that the rise of inflation could affect not only the amount of tourist's spending but may also cause a decrease in tourism demand as well.

Thailand's consumer price index (CPI) is acquired from <https://th.investing.com/>. The data on the website is provided by the Ministry of Commerce Thailand. Figure 22 indicates the fluctuation of the consumer price index throughout time.

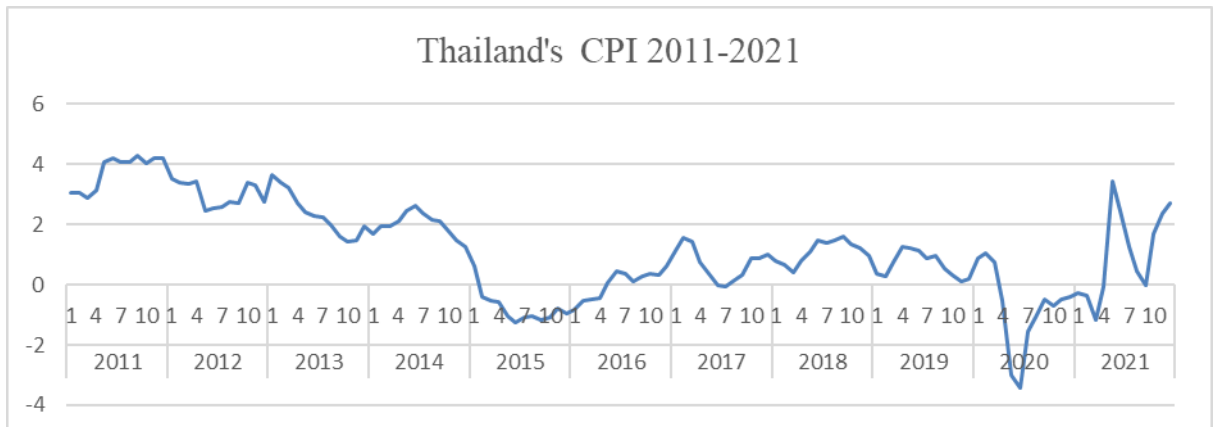


Figure 22. Thailand's consumer price index (CPI)

### 3.2.5.2 Jet fuel

Since one of the major airline operating costs is fuel cost, jet fuel is included as an explanatory variable. The monthly jet fuel price in Baht per Gallon is publicly available for download at <https://www.indexmundi.com/>. As illustrated in Figure 23, the price decline during the emergence of COVID-19 at the beginning of 2020, continuously increased over the course of 2021 and reached new highs in early 2022. The changes in jet fuel prices could be investigated to understand the correlation between air passenger demand and if it could potentially be used to assist in enhancing the prediction.

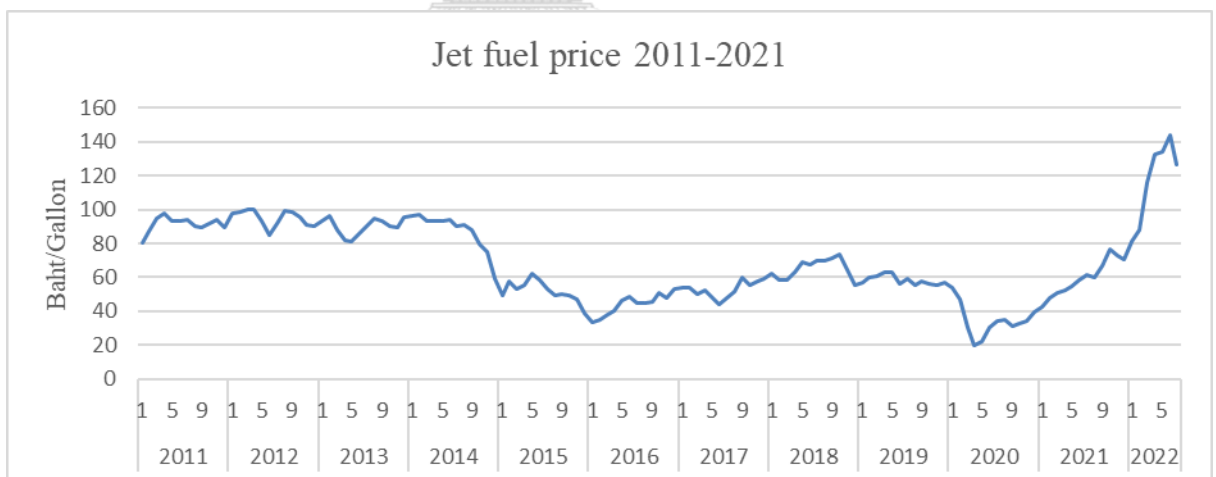


Figure 23. Jet fuel price

### 3.3 Preprocessing

Before raw data obtained from multiple sources are ready to be applied in a machine learning model, they need to be converted into a feasible format. Therefore, data preprocessing is an important procedure to prepare the input data. The historical volume of passengers comprises individual flights that arrived at the designated airport from specific originating countries. The airports are consolidated according to their locating region. The originating countries were disregarded; therefore, the total volume of arrival passengers is summed together according to the date and arrival region.

The data collected consists of all flights that arrived/departed from the airport, it may include Cargo flights and other flights with an absence of passengers, thus such data will be removed.

Log normalization is applied to standardize the volume of inbound passengers due to its high variance to reduce the impact of outliers as it could affect the performance of the model. The date of arrival is in object format, it will be converted into datetimes using the pandas function `to_datetime`. Useful features such as day, month, year, and day\_of\_week are extracted.

Concert and sporting events are incorporated as events. The event indicator is included in the model following the study of Catal et al. (2015) on introducing special days as a feature to enhance the ATM cash demand forecast by representing the specific days with a value of 0 or 1 in an additional column. Value 1 will represent the event held on the day and value 0 represent no event as shown in Table 8. Only the presence of events is considered as the exact number of attendees is unknown.

All features are measured on different scales. In order to avoid bias caused by such differences, they should be standardized. Scikit-Learn library provides `StandardScaler()` function which is applied to standardize the features by removing the mean and scaling to unit variance.

Date	Events_indicator
2011-04-23	0
2011-04-24	1
2011-04-25	1

Table 8. Events data representation

Features used to train the model can significantly impact performance in the sense of overfitting, accuracy, and training time. Adding irrelevant features can negatively impact model performance.

Due to the differences in the purpose of traveling and weather condition in each region, the linear correlation is calculated as of individual region to see which feature attributes has the most impact on the particular area. Queries selection using linear correlation was employed by Mavragani et al. (2020) to enhance COVID-19 case prediction and by Fu et al. (2022) to predict building energy consumption.

Correlation refers to the causal association between variables. The degree can interpret the strength of their relationship. Pearson correlation coefficient between 100 queries and Thailand's daily inbound passengers from 2011-2021 is calculated. The top 30 queries with the highest correlation are selected as explanatory variables. The three regions share similar top queries but differ in the magnitude of the correlation.

Studies by Yang et al. (2015) have led to a more profound understanding that web search activities executed at least one month prior to departure could enhance the tourism demand forecast. Cross-correlation is performed to find lag with the maximum correlation coefficient between the search volume and the volume of passengers using Matlab syntax `xcorr`.

The lags range from 0 to 4. The lag order indicates that visitors generally search for weather forecasts 4 months prior to the trip, then online travel services, flights, accommodations, and places to visit in the following months. The lag order confirms the nature of visitors in searching for various aspects of information during the planning (Yang et al., 2015). The lag orders of specific queries are different in the three regions (Central, Northern, and Southern).

Comparing Figures 24 and 25, the number of inbound passengers in the Central region in February shares a similar trend with the daily search volume on the query "Agoda Thailand" in January. The lag order calculated aligned with the graph observation, indicating that visitors search for "Agoda Thailand" 1 month prior to the trip.

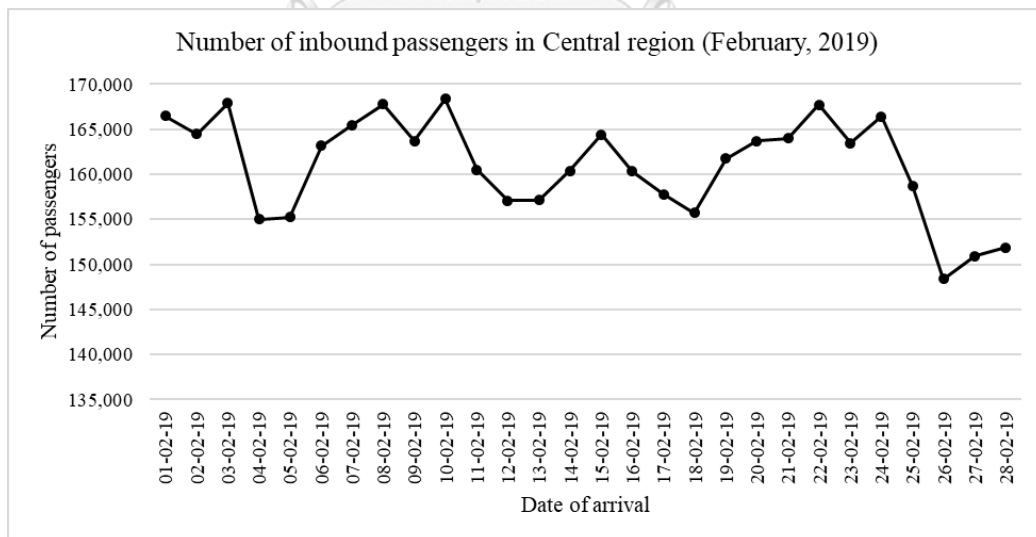


Figure 24. Number of inbound passengers in Central region (February 2019)



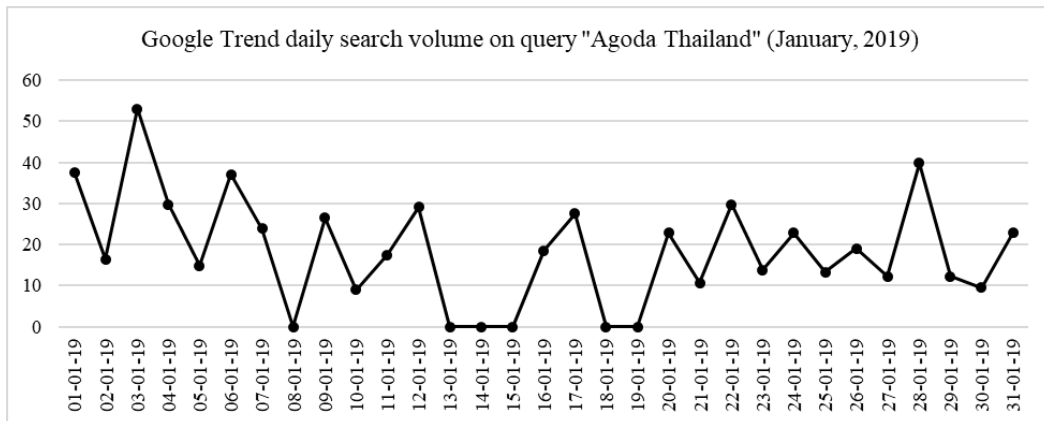


Figure 25. Daily search volume on query "Agoda Thailand" (January 2019)

Although the correlation between economic data and weather data is not as high as Google Trend queries, they will still be included as influencing variables in the assumption their combination may increase the prediction performance (Constantino et al., 2016; Varian et al., 2009; Vu et al., 2018). Figures 26, 27, and 28 illustrate the correlation of inbound passengers, weather data, CPI, and jet fuel prices in the Central, Northern, and Southern regions respectively. The events indicator is additionally included only in the Central region.

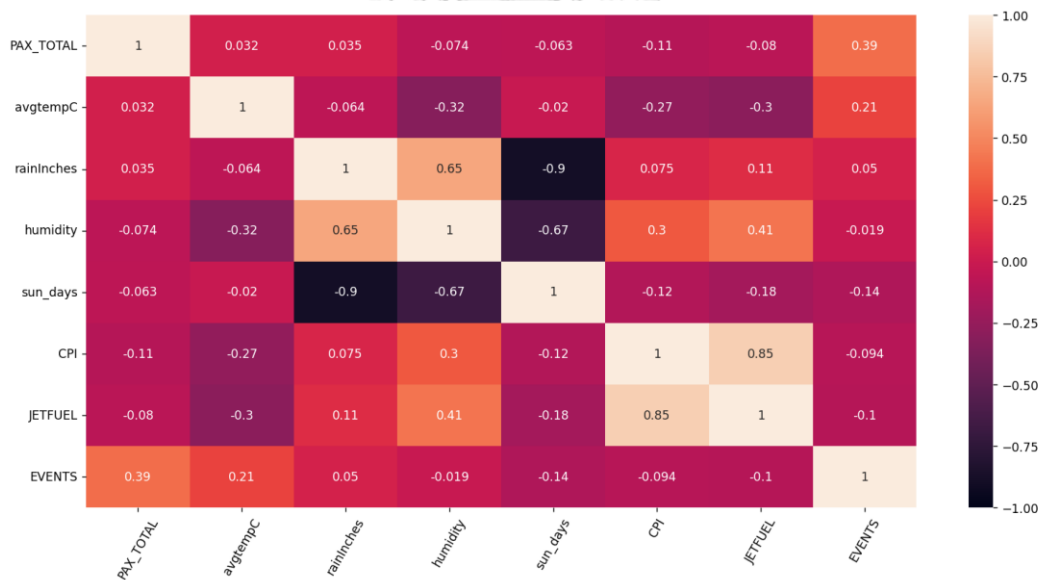


Figure 26. Correlation between Central inbound passengers, weather data, CPI, jet fuel prices, and events

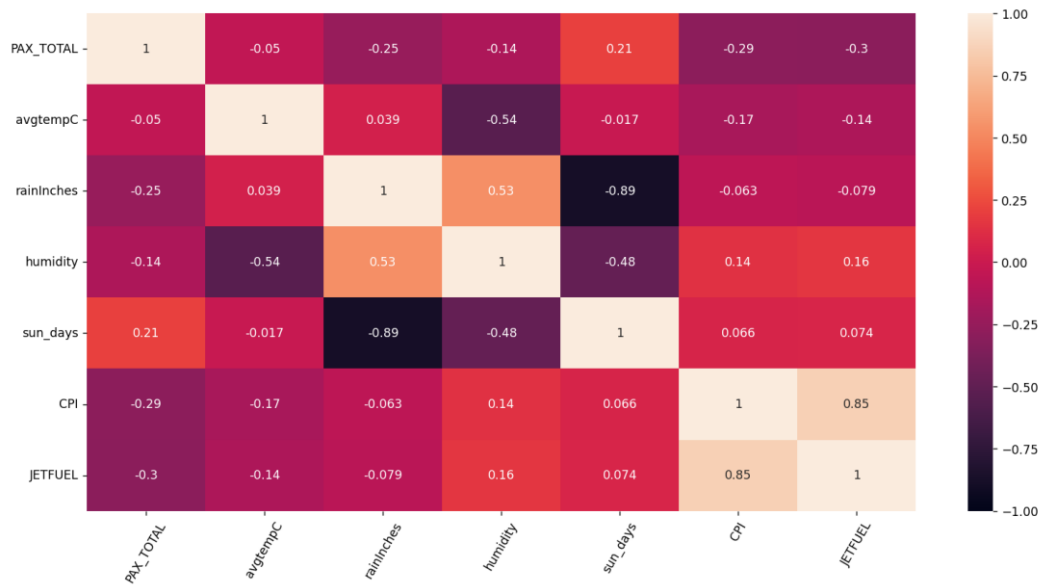


Figure 27. Correlation between Northern inbound passengers, weather data, CPI, and jet fuel prices



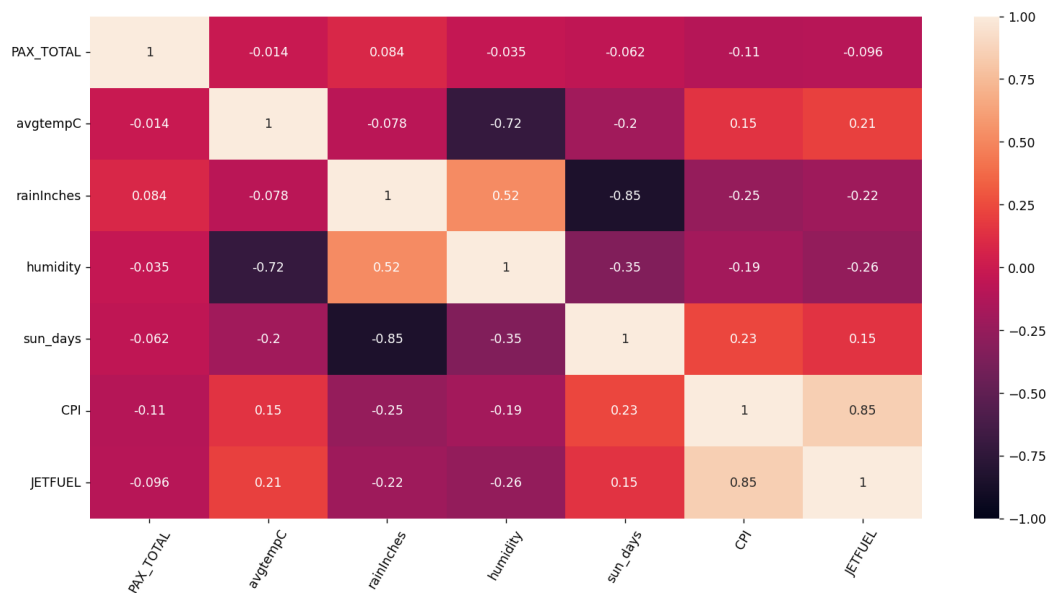


Figure 28. Correlation between Southern inbound passengers, weather data, CPI, and jet fuel prices

Overall, weather data did not show a high correlation with the inbound volume of passengers. However, the magnitude and direction of each feature are varied in different regions. *Sun\_days* regarding the average number of sunny days were excluded due to its extremely high negative correlation at over  $-0.85$  with *rainInches*, the attribute indicating the amount of rain in inches. Multicollinearity should be removed as it could reduce the model's performance.

The consumer price index (CPI) used to estimate the average differences in the prices of products consumed by households, shown to be negatively correlated with the inbound passengers in all regions, indicating that inflation could affect the travel budget and impact travel demand. Similarly, since jet fuel is considered to be the main operating cost, rises in jet fuel prices may escalate the airfare and impede the passenger from traveling, explaining the negative correlation.

Recursive Feature Elimination (RFE) (Guyon et al., 2002) is applied to evaluate the features based on how they affect the model's performance. The model complexity is reduced by removing features one at a time until an ideal number of features is realized. The combination of selected features from recursive feature elimination will be investigated to confirm their impact on the volume of passengers.

### 3.4 Regression method

Regression analysis allows the identification of relationships between a single independent variable and several dependent variables. A regression algorithm is utilized to create a model that will learn the connection between input and output data from labeled training data in order to make predictions from unseen data. Machine learning algorithms applied to develop the prediction in this research are Gradient Boosting Regression, Random Forest Regression, and Support Vector Regression, implemented in scikit-learn.

Gradient Boosting (Friedman, 2001) is one of the conventional algorithms for regression predictive modeling on account of its ability to handle multiple features and achieve high performance with minimum tuning. Dataset is divided into a subset to perform prediction and the model will learn from weaknesses in each iteration.

Random Forest (RF), introduced by Breiman (2001), is an ensemble machine learning algorithm constructed from a large number of decision trees, involving the evaluation of each input variable in the data aiming to select a split point. Unlike gradient boosting, the trees are independent. The average of the results allows the search for a more accurate predictive result while avoiding overfitting.

Support vector regression (SVR) (Drucker et al., 1996) is a supervised learning algorithm that is extended from a support vector machine. The objective of the algorithm is to search for a hyperplane in an n-dimensional space that classifies the data point. A kernel is generally used for finding a hyperplane. Different kernels are chosen upon the suitability of the dataset. The performance of SVR with Gaussian kernel and Polynomial kernel will be investigated.

The model settings and parameters were implemented in the same manner for using only historical inbound airline passengers and with the inclusion of external factors. The training and testing set is divided using Scikit-learn TimeSeriesSplit. Time series split ensures that train datasets are older than the test dataset. Data in the first 6 years is used as the training set and the rest as the testing set.

### 3.5 Evaluation Metrics

The performance of the model can be evaluated using many measurements. It is essential to quantify the performance of the model using appropriate measurements. In this chapter, the selected performance measure will be discussed followed by the forecasting result.

In general, metrics are chosen according to the types of predictions, classification, or regression. Widely used classification metrics in classification models are accuracy, precision, recall, and F1 score.

In contrast to classification, the objective of regression is not accuracy, and evaluating the model accordingly may develop overfitting.

Larsson et al. (2017) suggested that different types of forecasting errors can assist in identifying the strength and weaknesses of the forecast. A single metric may not be able to cover the entire perspective of the problem in consideration. Therefore, the evaluation metrics used to assess the model are MAE, MSE, and RMSE defined in (1), (2), and (3) using Scikit-Learn (Sklearn) and Numpy in Python.

Mean absolute error (MAE) is the average of the absolute differences between the actual value and the model's predicted value. Similarly, the Mean Square Error (MSE) is the square of MAE. The differences between actual and predicted values are squared to remove the sign so attention is paid to larger errors. Root Mean Squared Error (RMSE) is the square root of MSE used to measure the deviations between actual and predicted values. The lower the value for MAE, MSE, and RMSE, the better the prediction of the model.

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_i - \hat{y}_i| \quad (1)$$

$$MSE = \frac{1}{n} \sum_{t=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

Let  $n$  denotes the total number of observations in the data set. The actual and predicted value of the  $i$ th sample is represented by  $y_i$  and  $\hat{y}_i$  respectively.

## Chapter 4 Result and discussion

The experiments have been performed for three regions in Thailand (Central, Northern, and Southern) separately. The proposed methodology has been executed and evaluated using the historical volume of inbound passengers, Google Trend query, historical weather data, events indicator, jet fuel price, and consumer price index (CPI).

The prediction of passenger arrivals has been executed solely based on historical passenger arrivals and based on the addition of other variables. The period of historical passenger arrivals is from 2011 to 2019. Data after the COVID-19 outbreak will be investigated separately.

Additionally, events are included in the prediction model specifically for the Central Region according to the venue's location to investigate the prediction performance. Maximum correlation for events indicator is achieved with the lag of 4 days, indicating that passengers usually arrive at the destination 4 days before the events.

### 4.1 Kernel Selection

The performance of support vector regression with a Gaussian kernel and polynomial kernel is investigated using the same explanatory variables. As shown in Table 9, support vector regression with a Gaussian kernel outperformed the polynomial. The results reflect those of Claveria et al. (2015) who also concluded that support vector regression (SVR) with a Gaussian kernel significantly enhances the prediction of international tourism demand in Spain.

Kernel	MAE	MSE	RMSE
Gaussian	0.43175	0.19409	0.44056
Polynomial	0.57832	0.27354	0.60028

Table 9. Performance comparison between SVR with Gaussian kernel and SVR with polynomial kernel

The features selected by applying recursive feature elimination (RFE) for each region are shown in Table 10.

#### 4.2 Selected queries and lag order

Central	Northern	Southern
Jet fuel price	Jet fuel price	Jet fuel price
CPI	CPI	CPI
query: “Chiang Mai”	query: “Chiang Mai”	query: “Chiang Mai”
query: “Flight Thailand”	Rain	query: “Phi Phi Island”
query: “Phi Phi Island”	query: “Chiang Rai”	query: “Chiang Rai”
Humidity	Humidity	Humidity
query: “Thai Baht”	query: “Samui Island”	query: “Pattaya”
query: “Pattaya”	query: “Thai Baht”	query: “Thai Baht”
query: “Chatuchak Market”	query: “Khao Yai”	query: “Krabi”
query: “Krabi”	query: “Pattaya”	Sun_days

Table 10. Features selected by applying recursive feature elimination in each region

Some selected features are common in all regions. “Chiang Mai”, “Pattaya” and “Krabi” are among the top-rated tourist attractions in Thailand. Jet fuel prices and the consumer price index are considered important features that impact Thailand’s inbound passengers. As jet fuel is a major operating cost for airlines, an increase in jet fuel price could potentially raise airfare and affect passengers who are price sensitive. The assumption is supported by the negative correlation with the volume of passengers in the Central region as shown in Figure 29. The correlation for the Northern and Southern regions also follows the same trend.

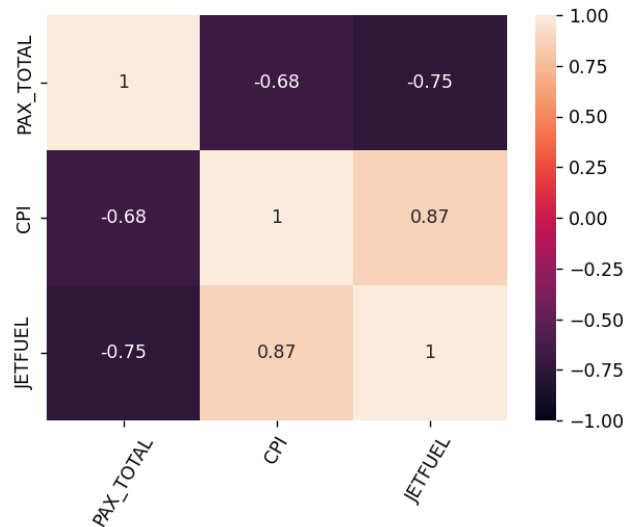


Figure 29. Correlation between CPI, jet fuel prices, and volume of passengers in Central region

The Google Trend data selected from performing recursive features elimination are shifted 0-7 months and Pearson correlation is used to obtain the appropriate lag that establishes the highest correlation. As shown in Table 11, although some features impact the volume of passengers in all regions, their lags are unique. Overall, passengers usually searched for flight information and places to visit 1-3 months prior to their arrival.

Google Trend Query	Lag Order (Central)	Lag Order (Northern)	Lag Order (Southern)
Chiang Mai	1	3	1
Chiang Rai	-	3	1
Samui Island	-	3	-
Thai Baht	1	1	1
Khao Yai	-	3	-
Pattaya	2	2	3
Chatuchak Market	2	-	-
Krabi	1	2	1
Phi Phi Island	1	-	1



<b>Flight Thailand</b>	3	-	-
------------------------	---	---	---

Table 11. Lag order of each query in different regions

### 4.3 Regression Results

This section will compare the performance of the prediction model with the inclusion of selected queries and illustrate the result from an addition of location-specific queries. The potential of using Google Trend queries in the prediction model after the COVID-19 pandemic will also be discussed.

#### 4.3.1 Performance of selected queries

Tables 12, 13, and 14 compare the performance of Gradient Boosting Regression (GB), Random Forest Regression (RF), and Support Vector Machine (SVR) with a Gaussian kernel by applying only historical data (HD) and the combination of additional features (AF).

With a unique combination of additional features selected from recursive feature elimination and shifted according to the lag order, the performance of all models has improved with the prediction error substantially decreasing. In particular, random forest regression outperformed Gradient Boosting Regressor and Support Vector Regressor.

Model		MAE	MSE	RMSE
GB	HD	0.43484	0.19298	0.43930
	AF	0.19404	0.04503	0.21221
RF	HD	0.49072	0.24554	0.49552
	AF	0.16688	0.03551	0.18844
SVR	HD	0.48353	0.24472	0.49469
	AF	0.43175	0.19409	0.44056

Table 12. Evaluation of regression model using 10 selected features from RFE (Central region)

Model		MAE	MSE	RMSE
GB	HD	0.62684	0.40586	0.63707
	AF	0.17131	0.04279	0.20686
RF	HD	0.74921	0.58027	0.76176
	AF	0.14400	0.02931	0.17121

<b>SVR</b>	<b>HD</b>	0.64098	0.42307	0.65044
	<b>AF</b>	0.20726	0.05820	0.24124

Table 13. Evaluation of regression model using 10 selected features from RFE (Northern region)

<b>Model</b>		<b>MAE</b>	<b>MSE</b>	<b>RMSE</b>
<b>GB</b>	<b>HD</b>	0.43986	0.20084	0.44815
	<b>AF</b>	0.21796	0.06128	0.24754
<b>RF</b>	<b>HD</b>	0.50803	0.26868	0.51834
	<b>AF</b>	0.24090	0.07803	0.27933
<b>SVR</b>	<b>HD</b>	0.41635	0.18968	0.43553
	<b>AF</b>	0.21319	0.05735	0.23947

Table 14. Evaluation of regression model using 10 selected features from RFE (Southern region)

The event feature was not selected from performing recursive feature elimination. The feature shows a relatively low correlation with the volume of a passenger in the Central region. The previous study stated that features with low correlation to the target were considered to be irrelevant independently, but they can be highly correlated with the presence of other features (Vu et al., 2018). The event feature is included in the model to examine the influence and the result shows a slight reduction in error as shown in Table 15.

#### 4.3.2 Location-specific queries

Model	Explanatory Variables	MAE	MSE	RMSE
<b>GB</b>	10 features from RFE	0.19404	0.04503	0.21221
	10 features from RFE + events	0.18397	0.04117	0.20289
	10 features from RFE + Events + location-specific queries	0.16473	0.03449	0.18572
	10 features from RFE + Events + location-specific queries + flood-related queries	0.14994	0.02797	0.16725
<b>RF</b>	10 features from RFE	0.16688	0.03551	0.18844
	10 features from RFE + events	0.16608	0.03503	0.18716
	10 features from RFE + Events + location-specific queries	0.15985	0.03337	0.18268
	10 features from RFE + Events + location-specific queries + flood-related queries	0.16106	0.03297	0.18158
<b>SVR</b>	10 features from RFE	0.43175	0.19409	0.44056
	10 features from RFE + events	0.23708	0.06706	0.25897
	10 features from RFE + Events + location-specific queries	0.23123	0.06384	0.25267
	10 features from RFE + Events + location-specific queries + flood-related queries	0.18462	0.04105	0.20262

Table 15. Comparing performance from the inclusion of additional features in the Central region

Aside from features that impact the overall regions, certain Google Trend queries affect only a particular region. “Chatuchak Market” is highly correlated with the volume of passengers in the Central region and proved to impact the prediction as it has been selected from recursive feature elimination. “Chatuchak Market”, located in Bangkok, is one of the world’s largest outdoor markets with over 15,000 stalls. From a wide range of products to street food, Chatuchak

market attracts over 200,000 visitors every week. The search query on “Chatuchak Market” shows potential in improving the prediction of passenger volume in the Central region.

Furthermore, several Google Trend queries are location-specific as well, thus their contribution is investigated individually. “Wat Pho” and “Wat Arun” are in the top 5 must-visit temples in Thailand, located nearby Chao Phraya River. Despite their low correlation, the inclusion of the terms shows a further reduction in error for the prediction of passengers in the Central region.

In addition, queries regarding the severe flooding during the 2011 monsoon were added to investigate the impact on passenger demand in the Central region. Due to its concentrated population, infrastructure, and issue with the drainage system, the flood crisis was more intense than in other regions. Table 15 illustrates that the queries had been shown to improve the overall prediction performance.

Among 30 queries selected from Pearson correlation, “Chiang Mai” and “Chiang Rai” which are located in the Northern region, show a high correlation in all regions and contribute to enhancing the performance of the prediction. One unique feature that impact the passenger arriving in the Northern is rain. Rain is negatively correlated to the volume of the passenger. Since the Northern region is famous for flower gardens and scenic mountain ranges, rain could undoubtedly impact the travel experience.

Apart from “Chiang Mai” and “Chiang Rai”, another query that is location-specific to the Northern region is “Doi Inthanon”. Doi Inthanon is the highest mountain in Thailand located in Chiang Mai. It is popular among tourists as it is known for its scenic viewpoints, nature trails, and many more. Although the term was excluded due to its low correlation, its inclusion was proved to reduce the prediction error in the Northern region as shown in Table 16.

Model	Explanatory Variables	MAE	MSE	RMSE
GB	10 features from RFE	0.62684	0.40586	0.63707
	10 features from RFE + location-specific queries	0.11033	0.01876	0.13696
RF	10 features from RFE	0.74921	0.58027	0.76176
	10 features from RFE + location-specific queries	0.09522	0.01509	0.12282
SVR	10 features from RFE	0.64098	0.42307	0.65044
	10 features from RFE + location-specific queries	0.10403	0.01610	0.12690

Table 16. Comparing performance from the inclusion of a location-specific feature in the Northern region

For the Southern region, low correlated queries, “Samui Island”, “Phuket”, “Ao Nang”, “Hat Yai”, and “Full Moon Party” are included in the model to investigate their contribution towards the volume of passengers. “Samui Island” and “Phuket” are famous islands located in the South of Thailand, known for beautiful beaches and hubs of scuba diving. “Ao Nang”, a resort town in Krabi, is also famous for snorkeling and diving. “Hat Yai” is the fourth largest city in the South, popular among tourists for shopping and visiting floating markets. Unlike other queries, “Full Moon Party” does not refer to a place but an event. “Full Moon Party” is considered to be one of the biggest events in Asia attracting visitors from all over the world. The event is held monthly when the moon is full at Phangan Island located in the South of Thailand. Although the event is scheduled monthly, the number of attendees may vary each month.

Table 17 demonstrates that the inclusion of the aforementioned queries can improve the performance of the prediction as overall errors are slightly reduced.

Model	Explanatory Variables	MAE	MSE	RMSE
GB	10 features from RFE	0.21796	0.06128	0.24754
	With addition of location-specific feature	0.20805	0.05664	0.23800
RF	10 features from RFE	0.24090	0.07803	0.27933
	With addition of location-specific feature	0.10933	0.01870	0.13674
SVR	10 features from RFE	0.21319	0.05735	0.23947
	With addition of location-specific feature	0.19175	0.04737	0.21764

Table 17. Comparing performance from the inclusion of a location-specific feature in the Southern region

Concluding that if a prediction is performed in each region separately, related queries selected should share the same geographic location with the concerned region.

Another interesting observation is the queries used by the searcher. Some queries are considered to be phonetic transcription, representing a spoken language in written form. For instance, the pronunciation of “temple” in the Thai language is “Wat”, thus most travelers would search for “Wat Arun” instead of “Arun temple”. In contrast, the pronunciation of “island” in the Thai language is “Koh” but the queries that are most searched for use “island” instead of “Koh” since “Phi Phi island” is among the top related queries from the initial keyword “travel Thailand”. The differences may depend on the searcher’s familiarity with the words. Travel sites could be using the word “Wat” more than “temple” and as

searchers came across the word often, they would start to search for more detail about the place.

#### 4.3.3 COVID-19 Pandemic

The result from the prediction using historical data from 2011 to 2021 without the inclusion of a dummy variable indicating the COVID-19 outbreak and other COVID-19-related queries shows a relatively large error in all regions as the pattern completely changed and the travel restriction is applied to the whole country. Table 18 illustrates that the dummy variable and COVID-19-related queries can potentially decrease the overall prediction error in the Central region.

Model	Explanatory Variables	MAE	MSE	RMSE
GB	10 features from RFE	0.43484	0.19298	0.43930
	With the addition of Dummy variable and COVID-19-related queries	0.25594	0.07956	0.28206
RF	10 features from RFE	0.49069	0.24560	0.49558
	With the addition of Dummy variable and COVID-19-related queries	0.25098	0.08030	0.28338
SVR	10 features from RFE	0.44141	0.20092	0.44824
	With the addition of Dummy variable and COVID-19-related queries	0.24827	0.07858	0.28032

Table 18. Performance from the inclusion of dummy variables and COVID-19-related queries for the Central region

In addition, the aforementioned Phuket sandbox would particularly impact the air passenger demand in the Southern region. Table 19 shows that the inclusion of the volume of search queries on the “Phuket sandbox” lowers the prediction error.

Although the queries had shown potential in improving the prediction, the error when the period covers the COVID-19 incident is relatively high, indicating there is still room for improvement.

Model		MAE	MSE	RMSE
GB	10 features from RFE	0.43986	0.20084	0.44815
	With the addition of “Phuket sandbox”	0.27244	0.09108	0.30179
RF	10 features from RFE	0.50827	0.26881	0.51847
	With the addition of “Phuket sandbox”	0.27331	0.09518	0.30851
SVR	10 features from RFE	0.43719	0.20381	0.45145
	With the addition of “Phuket sandbox”	0.26073	0.09201	0.30333

Table 19. Comparing performance from the inclusion of search volume on "Phuket sandbox" in the Southern region



## Chapter 5 Conclusion

### 5.1 Concluding the result

Following the work of Feng et al. (2019), initial keywords are selected, and related keywords are chosen accordingly. As for this research, inbound passengers are arriving from worldwide, and the search query can be in any language, related topics are also taken into consideration to overcome the language boundaries as suggested by Höpken et al. (2018).

The inbound volume of passengers used in this research is collected from 6 airports located in 3 regions of Thailand. Sharing similar weather conditions, the volume of passengers is combined according to their regions. The prediction is performed for each region separately to observe if disparate variables impact the performance differently.

By calculating the Pearson correlation between Google Trends queries and inbound passenger volume, queries with a correlation higher than 0.40 are selected. Recursive feature elimination is performed to identify the relevant subset of predictors that should be included in the model for predicting passenger volume in each region.

As has been previously reported in the literature on tourism prediction, distinct categories of the query were searched at various stages of trip planning (Yang et al., 2015). The lag order calculated by cross-correlation provided the suitable period of search volume to be utilized in the prediction model. For instance, the maximum correlation between the search query “Flight Thailand” and passenger volume in the Central region is acquired with a lag of 3. Indicating that travelers usually searched for information on flights 3 months prior to their arrival. Thus, the current search volume on “Flight Thailand” can be incorporated in predicting the arrival of passengers in the Central region 3 months ahead.

With the inclusion of a certain combination of Google Trend data, weather data, events indicators, consumer index prices, and jet fuel prices, the performance of the prediction had shown observable improvement as forecasting errors decreased. However, some unique variables had been shown to impact the particular region exclusively. Individual Google queries selected do not reflect the volume of passengers traveling to Thailand overall but in a particular region.

As some scholars suggested, feature filtering does not necessarily improve predictive performance. The correlation between the features and the target might be relatively low but dropping them might as well decrease the performance (Constantino et al., 2016; Varian et al., 2009; Vu et al., 2018). Although feature selection is performed in this research, not only the features with the highest correlation were included. As discussed by Höpken et al. (2018), high correlation does not designate predictive power. Thus, queries with lower correlation are added to investigate the influences.

The result from recursive feature elimination indicates that search volume for the “Chatuchak market” impacts the volume of passengers, particularly in the Central region. Thus, location-specific queries with low correlation are further investigated. The forecasting error drops with the inclusion of “Wat Pho”. The same results are obtained when location-specific queries are added in correspondence with the regions regardless of their relatively low correlation. The



results were aligned with the prior research as the best prediction performance did not include one variable exclusively but a combination of relevant features.

As observed from Table 10, the Google Trend queries on “Chiang Mai” and “Pattaya,” are highly correlated to the volume of passengers in all regions. Both are cities ranked in the top 5 destinations visited by international tourists. Travelers who plan to visit Thailand may initially explore multiple popular destinations before making a final decision.

Another highly correlated Google Trend query to the overall volume of passengers is “Thai Baht”. Lag time indicates that travelers are searching for Thai currency 1 month prior to the trip as they may start to check the exchange rate. If the originating country is taken into consideration, the query used might be more specific, for instance, “USD to Baht”.

Consumer price index and jet fuel price were selected when recursive feature selection is performed. They improved the performance of the prediction in all regions but did not outperform the combination of features as listed in the table. However, the result can conclude they also relatively impact the volume of passengers in all regions of Thailand.

Not only that jet fuel may contribute to increases in airfare and affect price-sensitive passengers, but it can eventually change the airline operation. The airline could reevaluate route profitability and consider reducing the flight frequency or stopping operating a certain route. Such action could potentially impact the air travel demand and consequently tourism.

Weather data, consumer price index, jet fuel, and Google Trend queries were proven to impact Thailand’s inbound passenger volume. A selection of an appropriate set of variables can assist in demand prediction to support the airline in decision-making and future planning.

The addition of flood-related queries that emerged in mid-2011 proved that rising related topics and related queries could assist in demand prediction and explain the shift in demand during a short period whereas considering only top related topics and queries may have been missed.

The COVID-19 pandemic disrupted the aviation industry, and the air passenger demand prediction shows relatively large errors may be due to dissimilar trends. However, IATA (2022) recently indicates that the number of international travelers is expected to improve to 101% compared to 2019 levels in 2025. Thus, as more data after the pandemic accumulates, the model could be retrained to make a better forecast.

The proposed approach is not limited to passenger demand prediction but may potentially be applied to tourism and other related businesses that share a similar characteristic.

## 5.2 Future work

The unavailability of reliable sources for event data is a major constraint in this research as the detail should include the size of the venues, especially, the number of attendees. It is difficult to justify whether the events are attracting fans globally or the attendees are the domestic population. Moreover, social media data can be used as a tool to assist in measuring the popularity of the events

similar to the previous research that utilized online reviews to gain useful insights into the trend related to the local attractions (Varian et al., 2009). However, the challenge lies in the fact that although the events are trending on social media, it does not practically mean they will fly to attend.

Additionally, other platforms of social networks such as Facebook and Twitter can be used to overcome the limitation of using search queries in the absence of sentiment.

This research did not take the originating countries of the arrival passengers into account. The econometric features such as exchange rates could be added to enhance the performance, but the limitation is the originating country must be specific due to different currency exchange rates. If the originating country is taken into consideration, it is also suggested that the choice of search engine used must be selected upon the users' preference. Although Google accounts for the biggest search engine market, previous research had shown that the choice of search engine used varies in each country. For instance, considering China as the originating country, Baidu is the largest search engine in China with over 200 million active users a day. Google search engine may not reflect their preferences as effectively as Baidu. Moreover, according to previous research stating that the explanatory variables are country-specific, top, or rising related queries from searchers worldwide may not be effective as specific terms important to a certain country could be excluded from the ranking.

## REFERENCES

- Adjallah, J. (2022). *Brussels Airlines To Transport 25k Festival-Goers to Tomorrowland*. <https://travelradar.aero/brussels-airlines-to-transport-25k-festival-goers-to-tomorrowland/>
- Amusa, L. B., et al. (2022). Modeling covid-19 incidence with google trends. *Frontiers in Research Metrics and Analytics*, 7.
- An, B., et al. (2016). *MAP: Frequency-Based Maximization of Airline Profits based on an Ensemble Forecasting Approach* In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16),
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32.
- Catal, C., et al. (2015). Improvement of demand forecasting models with special days. *Procedia Computer Science*, 59, 262–267.
- Choi, S., et al. (2016). *Prediction of weather-induced airline delays based on machine learning algorithms* IEEE/AIAA 35th Digital Avionics Systems Conference (DASC),
- Claveria, O., et al. (2015). Regional forecasting with support vector regressions: The case of Spain. *SSRN Electronic Journal*.
- Constantino, H. A., et al. (2016). Tourism demand modeling and forecasting with artificial neural network models: The Mozambique Case Study. *Tékhné*, 14(2), 113–124.
- Drucker, H., et al. (1996). *Support vector regression machines* 9th International Conference on Neural Information Processing Systems,
- Eglitis, L. (2020). *International tourism: the most popular countries*. <https://www.worlddata.info/tourism.php>
- Fahad, A., et al. (2013). Forecasting air travel demand of Kuwait: A comparison study by using regression vs. artificial intelligence. *Journal of Engineering Research*, 1, 113-114.
- Feng, Y., et al. (2019). Forecasting the number of inbound tourists with google trends. *Procedia Computer Science*, 162, 628–633.
- Fesenmaier, D. R., et al. (2010). An analysis of search engine use for travel planning. *Information and Communication Technologies in Touris*, 381–392.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232.
- Fu, C., et al. (2022). Using google trends as a proxy for occupant behavior to predict building energy consumption. *Applied Energy*, 310.
- Ghalekhondabi, I., et al. (2019). A review of demand forecasting models and methodological developments within tourism and passenger transportation industry. *Journal of Tourism Futures*, 5(1), 75–93.
- Guyon, I., et al. (2002). Gene Selection for Cancer Classification using Support Vector Machines. 46, 389-422.
- Höpken, W., et al. (2018). Google Trends data for analysing tourists' online search behaviour and improving demand forecasting: The case of Åre, Sweden. *Information Technology and Tourism*, 21(1), 45–62.
- IATA. (2022). *Air Passenger Numbers to Recover in 2024*. <https://www.iata.org/en/pressroom/2022-releases/2022-03-01-01/>
- Kim, J., et al. (2017). Role of tourism price in attracting international tourists: The case

- of Japanese inbound tourism from South Korea. *The Journal of Destination Marketing & Management (JDMM)*, 6(1), 76–83.
- Kort, R. E. d. (2017). Forecasting tourism demand through search queries and machine learning.
- Larsson, F., et al. (2017). An Analysis of Passenger Demand Forecast Evaluation Methods.
- Li, H., et al. (2020). Forecasting tourism demand with Multisource Big Data. *Annals of Tourism Research*, 83, 102912.
- Liu, J., et al. (2017). Personalized Air Travel Prediction: A Multifactor Perspective. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(3), 30.
- Long, C. L., et al. (2021). Air passenger forecasting using neural granger causal google trend queries. *Journal of Air Transport Management*, 95, 102083.
- Mao, L., et al. (2015). Modeling Monthly flows of global air travel passengers: An open-access data resource. *Journal of Transport Geography*, 48, 52–60.
- Mavragani, A., et al. (2020). Covid-19 predictability in the United States using google trends time series. *Scientific Reports*, 10(1).
- Morley, C. L. (1994). The use of CPI for tourism prices in demand modeling. *Tourism Management*, 15(5), 342–346.
- Nada, K., et al. (2012). Modelling Seasonal Variation in Tourism Flows with Climate Variables. *Tourism Analysis*, 121–137.
- National Statistic Office Ministry of Digital Economy and Society. (2021). Statistical yearbook Thailand 2021.
- National Statistical Office of Thailand. (2022). *17 Tourism and Sports Branch*. <http://statbbi.nso.go.th/staticreport/page/sector/en/17.aspx>
- Olmsted, L. (2022). *Sports Travel—How To Attend The Biggest Sports Events In The World*. <https://www.forbes.com/sites/larryolmsted/2022/08/09/sports-travelhow-to-attend-the-biggest-sports-events-in-the-world/?sh=6ee488409ba0>
- Pektaş, Ş. Y. (2020). The Evaluation of Tourism in Turkey in Terms of Inflation. *Journal of Tourismology*, 6(1), 125-146. <https://doi.org/10.26650/jot.2020.6.1.0012>
- Prideaux, B., et al. (2006). Events in Indonesia: Exploring the limits to formal tourism trends forecasting methods in complex crisis situations. *Crisis management in tourism*, 353–374.
- Rashad, A. S. (2022). The power of travel search data in forecasting the tourism demand in Dubai. *Forecasting*, vol. 4(no. 3), 674–684.
- Riedel, S., et al. (2003, 10 - 12 July 2003). *Adaptive Mechanisms in an Airline Ticket Demand Forecasting System* EUNITE'2003 Conference: European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems, Oulu, Finland.
- Scott, D., et al. (2010). Weather and climate information for tourism. *Procedia Environmental Sciences*. 1, 146–183.
- Uzama, A. (2009). Marketing Japan's travel and Tourism Industry to international tourists. *SSRN Electronic Journal*.
- Varian, H. R., et al. (2009). Predicting the present with google trends. *SSRN Electronic Journal*.
- Vu, V. H., et al. (2018). *An airfare prediction model for developing markets* International Conference on Information Networking (ICOIN),

- Wang, T., et al. (2019). *A framework for airfare price prediction: A machine learning approach* IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI),
- World Tourism Organization. (2021). *International Tourism Highlights, 2020 Edition*. <https://doi.org/https://doi.org/10.18111/9789284422456>
- Xiong, H., et al. (2022). A novel approach to Air Passenger Index Prediction: Based on mutual information principle and support vector regression blended model. *SAGE Open*, 12(1), 215824402110711.
- Yang, X., et al. (2015). Forecasting Chinese tourist volume with Search Engine Data. *Tourism Management*, 46, 386–397.
- Yuan, H. (2014). *A user behavior-based ticket sales prediction using data mining tools: An empirical study in an OTA company* 11th International Conference on Service Systems and Service Management (ICSSSM), Beijing.





จุฬาลงกรณ์มหาวิทยาลัย  
**CHULALONGKORN UNIVERSITY**

## VITA

**NAME** Sutthiya Lertyongphati  
**DATE OF BIRTH** 21 May 1994



จุฬาลงกรณ์มหาวิทยาลัย  
**CHULALONGKORN UNIVERSITY**