Forecasting Thailand inbound tourist flow association for tourism demand

Danhua Jiang¹, Jianxu Liu²,³, Jirakom Sirisrisakulchai¹,³,⁴ and Songsak Sriboonchitta¹,³

¹Faculty of Economics, Chiang Mai University, Chiang Mai 50200, Thailand
²Faculty of Economics, Shandong University of Finance and Economics, Jinan, China
³Puey Ungphakorn Center of Excellence in Economics, Chiang Mai University, Chiang Mai 50200, Thailand
⁴Corresponding author’s e-mail: sirisrisakulchai@hotmail.com

Abstract. The purpose of this paper is to study the tourism demand of Thailand from six major source countries (CN, JP, KO, MAS, SG and LA) in two regions of Asia (East Asia and Southeast Asia). The data is from 1997q1 to 2020q1, which can also provide some advice for the recovery of Thailand tourism post COVID-19. Compared with previous studies, we consider the dependence of tourist flow in the same region. Precisely, we estimated the impact of tourism flows between pairwise countries in each region. This also provides a complement to the subsequent research on the structure of tourism dependence for tourism demand of Thailand. The Copula-based approaches are increasingly being used for tourism demand, we apply the copula-ARDL (ECM) framework to forecast tourist arrivals from two regions. Using the characteristics of correlation matrix in the trivariate Gaussian copula model, we have innovatively performed unconstrained parametric optimization of these two models to ensure the accuracy of the forecasting.

1. Introduction

The World Travel and Tourism Council’s (WTTC 2019) shows that tourism accounted for 10.4% of global GDP, creating jobs, boosting exports and creating prosperity around the world, with the fastest growth especially in developing countries. Thailand is a classical developing country known for its tourism, as one of the most popular tourist destinations in the world, modeling and forecasting of tourism demand in Thailand has been a hot topic. According to data from Thomson Reuters through the past 13 years, the top ten countries for inbound tourists from Thailand are China, Malaysia, Japan, South Korea, UK, Singapore, US, India, Laos, and Russia. In the case of Thailand, the cumulative number of Asian tourists reached the top, it has made a significant contribution to the development of Thailand's tourism industry. However, many previous studies have been limited to study the impact of mainstream factors on inbound tourism to Thailand, such as income of the tourist-generating (origin country), transportation costs and tourism price of goods and services [1]. But they did not realize that the dependence between tourism demand was a point to be considered, which may reduce the accuracy of the traditional model to forecast the tourism demand.

According to the Pearson correlation, Copula performs better than traditional methods in forecasting because of its flexible characteristics. And it can break the constraints of normal distribution and linear hypothesis, estimate the dependence of the growth rate of the tourist arrivals of two countries to Thailand [2]. L. Zhu and J. Zhang [3] also proposed to use the Bernstein Copula model of
high-dimensional correlation to solve the problem of arbitrary dependence between dependent variables and covariates. The empirical results show that the Copula model is more effective in forecasting inbound tourism demand and they also emphasize the importance of dependence structure to tourism demand.

Specifically, we use the copula approach to forecast tourist arrivals from the six main source countries in Asia based on the previous data. According to the regional location and economic situation in Asia, we screened out six countries (China, Japan, South Korea, Malaysia, Singapore and Laos) to form two groups to study the dependence of pairwise tourist flows.

2. Literature review
In recent years, methods of forecasting tourism demand can be divided into four categories [4]. The first three are time series models, econometric models, and AI-based models, which belong to quantitative methods [5]. The last category is the judgmental method, which can be used for forecasting qualitative and quantitative [6]. Amongst these methods, time series and econometric models are the most popularly used. However, the forecasting performance of the time series model is not good, and the explanatory power is limited. Therefore, more and more papers have started to use econometric models to study tourism demand. Although artificial intelligence and big data (AIBD) have the advantages of timely forecasting tourism demand and overcoming the lag of traditional forecasting methods [7], there are limitations in theory and technology [5]. In recent studies, the combined forecasting of the combined model and the hybrid model is often used because the accuracy cannot be lower than their constituent forecasts [8]. Since the complexity of tourism demand, a single approach cannot satisfy many situations at the same time. So far, forecasting methods are improving, and the accuracy of forecasting has been improved. Copula model, which is often used in finance, insurance and economic research, has been extended to tourism demand forecasting. The judgment method is more often combined with the above models to improve the accuracy of the quantitative forecasting model.

Lim [9] summarized 100 papers to analyze the variables of tourism demand and found that the most frequently cited explanatory variable was income level, followed by relative price and transportation cost. In addition, alternative price also has a great effect on tourism demand [10], while international tourism demand might be affected by some other qualitative factors.

In addition to the well-known major variables affecting tourism demand, more and more articles focus on the dependence on tourism demand variables. By using copula-GARCH models or panel data techniques investigates dependence between tourism demand and exchange rate [11]. The empirical results show that the exchange rate fluctuations have different impacts on tourism demand in each country. And using panel data techniques to explore the dependence between GDP and tourism receipts growth rates [12], it was found that basically there was a significant, asymmetric and positive correlation. We mentioned above that most extant literatures either only considered the linear relationship in normal distribution between tourist flows or ignored the dependency structure between flows from different source markets. This means that the implicit assumption of these papers is that tourism flows are independent of each other. Once tourism flows are correlated, the lack of calculation for dependence structure will inevitably reduce the accuracy for tourism demand forecasting. Athanasopoulos and De Silva [13] account tourist flows correlation into tourist arrivals in Australia and New Zealand, which could improve the accuracy of prediction. Similarly, L. Zhu et al. [14] also demonstrated that the combined tourism demand model with optimal dependency structure had a better fitness value. This shows that dependency structure is an important problem to be considered in the tourism demand model.

3. Methodology

3.1. Copula model
Sklar’s Theorem (Sklar 1959) can write any joint distribution in the form of copula, which provides a more general framework for copulas when describing joint distribution and set the foundation for
copulas. Consider random variables $X$ with a $n$–dimensional cumulative distribution function c.d.f. $F(\cdot)$, and margins $F_1, F_2, \ldots, F_n$, then there exists a copula function $C(\cdot)$:

$$F(x_1, \ldots, x_n) = C[F_1(x_1), F_2(x_2), \ldots, F_n(x_n)],$$  

(1)

where $x_i$ is the observation of $X_i$.

In view of Sklar’s Theorem, a unique copula multivariate distribution function can be used to represent the joint distribution of tourism flows in different regions $(Q_{l,t}, i = 1, 2, \ldots, n)$ as follow:

$$F(q_{1,t}, q_{2,t}, \ldots, q_{n,t}) = C[F_{q1}(q_{1,t}), F_{q2}(q_{2,t}), \ldots, F_{qn}(q_{n,t}); \Theta],$$  

(2)

where $F(\cdot)$ is the joint distribution of $Q_{l,t}$; $F_{q1}(\cdot)$ is the marginal distribution of tourist arrivals $Q_{1,t}$; $\Theta$ denotes the parameter vector of the copula function, and $C(\cdot; \Theta)$ represents the information captured by the $\Theta$ that depends on the dependence among the of the $Q_{l,t}$'s.

3.2. Model Specification for tourism demand

A general model for studying tourism demand can be written in the following form,

$$\ln Q_{l,t} = D_t(\mathbf{X}_{l,t}), \quad i = 1, 2, \ldots, n ,$$  

(3)

where $Q_{l,t}$ is the tourist arrivals from origin country $i \in (1, 2, \ldots, n)$ in time $t \in (1, 2, \ldots, T)$; $D_t$ is the tourism demand function of the $i$-th flow, for which the variable of tourist arrivals to destination country; $\mathbf{X}_{l,t}$ is the vector of explanatory variables for origin country $i$ at time $t$.

The ARDL and ECM models are often used in forecasting tourism demand, and their performance is better. Moreover, ARDL model can not only estimate the influence of lagging factors but integrates the influence of lagging demand variables. On this basis, ECM model further determines the long-term relationship between tourism demand and influencing factors. The ARDL model of equation (3) is

$$\ln Q_{l,t} = \alpha_{l,0} + \sum_{s=1}^{l_1} \alpha_{l,s} \ln Q_{l,t-s} + \sum_{s=0}^{l_2} \beta_{Y_{l,s}} \ln Y_{l,t-s} + \sum_{s=0}^{l_3} \beta_{P_{l,s}} \ln P_{l,t-s} + \sum_{s=0}^{l_4} \beta_{OP_{l,s}} \ln OP_{l,t-s} + \text{Dummies} + e_{l,t} \sim \mathcal{N}(0, \sigma^2),$$  

(4)

where $Y_{l,t}$ is the income level of the origin country $i$ at time $t$, measured by the logarithm of GDP. $OP_{l,t}$ is the logarithm of international crude oil price (West Texas Intermediate oil price) is the proxy for transportation cost. $\beta_{s,t}$ is the time series parameter. $l_i$ is the number of lags of tourism demand and the independent variables, $e_{l,t}$ is the error term. $P_{l,t}$ is tourism relative price, adjusted by the relevant exchange rates as follows:

$$P_{l,t} = \ln \frac{\text{CPI}_{i, t}}{\text{CPI}_{j, t}}$$  

(5)

$EX_i$, $EX_j$, $\text{CPI}_i$, $\text{CPI}_j$ are the exchange rate and consumer price index (base year 2010) of different origin country $i$ and destination country $j$.

Engle and Granger (1987) showed that the error correction model (ECM) can link the change of a variable with the past equilibrium error. When a long-run random trend occurs, it is called cointegration, and they are closely related. The model provides a dynamic method that can combine long-run equilibrium and short-run non-equilibrium adjustment processes simultaneously, thereby avoiding economic variables from deviating from their long-run equilibrium. After generating the error correction term, ECM to be showed is expressed as

$$\Delta Q_{l,t} = \pi_{l,0} + \pi_{Y_{l,t}} \Delta Y_{l,t} + \pi_{P_{l,t}} \Delta P_{l,t} + \pi_{OP_{l,t}} \Delta OP_{l,t} + \text{Dummies} + \pi_{e} \tilde{e}_{l,t-1} + v_{l,t},$$  

(6)

with symbol $\Delta$ performing first-difference for each specific indicator respectively, $\tilde{e}_{l,t-1}$ and $v_{l,t}$ are the error term. $\tilde{e}_{l,t-1}$ is given by which is:

$$\tilde{e}_{l,t-1} = Q_{l,t-1} - \alpha_{l,0} - \beta_{Y_{l,t-1}} Y_{l,t-1} - \beta_{P_{l,t-1}} P_{l,t-1} - \beta_{OP_{l,t}} OP_{l,t-1} - \text{Dummies}.$$  

(7)

Based on the method of probability estimation, Eq. (6) and Eq. (7) are estimated simultaneously. The log likelihood function of $n$ tourism flows is:

$$L = \sum_{t=1}^{T} \left( \ln f_{1,t} + \ln f_{2,t} + \cdots + \ln f_{n,t} \right) + \sum_{t=1}^{T} \ln c[F_{1,t}(q_{1,t}), F_{2,t}(q_{2,t}), \ldots, F_{n,t}(q_{n,t}); \Theta],$$  

(8)

where $f_{i,t}$ is $F_{i,t}$ marginal distribution density, $\Theta$ shows tourism flows are obtained by the association parameter, so that information about the dependency between $Q_{1,t}$ and $Q_{2,t}$ captured. And $c(\cdot)$ is the probability density function given by the copula distribution function $C(\cdot)$, as follows:

3
\[
c(\cdot) = \frac{\partial^n c}{\partial F_{1,t} \partial F_{2,t} \cdots \partial F_{n,t}}. \tag{9}
\]

But it is difficult to force that correlation matrix for Gaussian copula be positive semi-definite certain fixed matrix. In order to satisfy this condition, we applied the unconstrained parameterization of the variance-covariance matrices to ensure the positive definiteness. And ensuring estimation will not be constrained while leaving the positive definite value. Moreover, the parameterized estimated variance-covariance matrix is easy to work and ensures the stability of the entire matrix. Following Pinheiro and Rapisarda, we make sure that \( R \) remains the symmetric positive definite \( 3 \times 3 \) matrix for maximizing the likelihood, \( R \) be decomposed as

\[
R = LL^T,
\]

where \( L \) is a lower triangular matrix with nonnegative diagonal elements, it has been parameterized. So, the correlation matrix \( R \) is expanded as follows:

\[
R = \begin{bmatrix}
\cos\theta_{12} & \cos\theta_{13} & 0 \\
\cos\theta_{12} & \cos\theta_{13} & \sin\theta_{12}\sin\theta_{13} \\
0 & \sin\theta_{12}\sin\theta_{13} & \sin\theta_{23}\sin\theta_{13}
\end{bmatrix}
= \begin{bmatrix}
1 & \rho_{12} & \rho_{13} \\
\rho_{12} & 1 & \rho_{23} \\
\rho_{13} & \rho_{23} & 1
\end{bmatrix}
\]

where correlation coefficients \( \rho_{ij} \) can be expressed as

\[
\rho_{12} = f(\theta_{12}) = \cos\theta_{12}, \rho_{13} = f(\theta_{13}) = \cos\theta_{13}, \rho_{23} = f(\theta_{12}, \theta_{13}, \theta_{23}) = \cos\theta_{12}\cos\theta_{13} + \sin\theta_{12}\sin\theta_{23}\sin\theta_{13}.
\]

4. Empirical results

4.1. Data description

We divide the estimated sample of this paper into estimation period (1997q1-2013q4) and verification period (2014q1-2020q1). Due to the COVID-19, we just use the data until 2020q1. The data of all variables obtained from Thomson Reuters, except Laos CPI 1997q1-1999q4, this part is calculated based on the growth rate of Laos CPI from CEIC.

The dummy variables include D1997 in the Asian financial crisis in 1997, D2003 in the SARS epidemic in 2003, D2008 in the global financial crisis in 2008, and D2010 in Thailand’s anti-government demonstrations in 2010.

4.2. Model estimation

In this study, the Augmented Dickey Fuller test method is used to test the stationary of variables \((Q_{it}, Y_{it}, P_{it}, OP_{it})\) in the international tourism demand model. Table 1 shows that the first-order difference of each variable is stationary, because each related variable exceeds the critical value of −2.58 under the significance level of 10%, which is I(1).

| Variable | East Asia | Southeast Asia |
|----------|-----------|----------------|
|         | CN       | JP  | KR  | MAS  | SG  | LA  |
| Qit      | -0.524   | -1.142 | -1.200 | 0.332 | -1.085 | -1.952 |
| Yit      | -1.561   | -2.131 | -0.727 | -0.327 | -0.116 | -0.856 |
| Pit      | -1.628   | -1.271 | -2.508 | 0.161 | -3.893*** | -2.489 |
| OP1t     | -1.664   |        |        |       |       |      |
|          | Augmented Dickey–Fuller statistic (level) |
| Qit      | -14.220*** | -8.660*** | -6.876*** | -7.456*** | -13.615*** | -8.566*** |
| Yit      | -2.423    | -4.261*** | -11.374*** | -7.500*** | -7.736*** | -2.592* |
| Pit      | -9.663*** | -8.614*** | -13.225*** | -8.255*** | -8.686*** | -8.717*** |
| OP1t     | -7.906*** |        |        |       |       |      |
Table 2 shows which country can pass the boundary test when the ARDL model is used, and the number of lags of the independent variables of each country which are determined by the Schwarz criterion (SBIC). As shown in the Table 2, Malaysia, Singapore and Laos (Group 2) have passed the boundary test, indicating that there is a long-term cointegration relationship and lagged variable is important for models of these countries because they are significant. Therefore, their tourism demand can be estimated with ARDL copula model. In contrast, China, Japan, and South Korea (Group 1) failed the boundary test, the Johansen test cointegrate was used. In Table 3, there is at least one cointegration relationship in all variables of Group 1, so we use ECM copula model that allows the existence of cointegration to estimate short-term elasticity of tourism demand.

### Table 2. Bounds test for cointegration and ARDL specification.

|                        | Best Statistical significance | Critical values of bound test |
|------------------------|------------------------------|------------------------------|
|                        | F-statistic | ARDL (p1, p2, p3, p4) | L  | 0.1 (90%) | 0.05 (95%) | U  |
| East Asia              |             |                          |    |           |             |    |
| CN                     | 3.553       | -                        | 2.37 | 3.2 | 2.79 | 3.67 |
| JP                     | 3.361       | -                        |    |       |             |    |
| KR                     | 3.449       | -                        |    |       |             |    |
| Southeast Asia         |             |                          |    |           |             |    |
| MAS                    | 4.819       | (4,0,0,0)                |    |       |             |    |
| SG                     | 4.011       | (4,0,0,3)                |    |       |             |    |
| LA                     | 4.279       | (2,0,0,0)                |    |       |             |    |

*The order of lagging items in brackets is $Q_{it}$, $Y_{it}$, $P_{it}$, $OP_t$. L and U are lower bound and upper bound.

### Table 3. Johansen cointegration test.

| Hypothesized No. of CE(s) | CN | JP | KR |
|---------------------------|----|----|----|
| Unrestricted cointegration rank test (Trace) | Trace Statistic | P | Trace Statistic | P | Trace Statistic | P |
| None                      | 46.125 | 0.072 | 30.325 | 0.008 | 28.311 | 0.015 |
| At most 1                 | 21.338 | 0.337 | 10.722 | 0.091 | 11.576 | 0.066 |
| At most 2                 | 4.798 | 0.830 | 0.885 | 0.401 | 3.303 | 0.082 |
| Unrestricted Cointegration Rank Test (Maximum Eigenvalue) | Trace Statistic | P | Trace Statistic | P | Trace Statistic | P |
| None                      | 24.786 | 0.110 | 19.603 | 0.027 | 16.735 | 0.072 |
| At most 1                 | 16.540 | 0.195 | 9.837 | 0.087 | 8.272 | 0.158 |
| At most 2                 | 3.986 | 0.861 | 0.885 | 0.401 | 3.303 | 0.082 |

### 4.3. Forecasting

Through the previous estimation, we know that different lagged variables ($Q_{it}$, $Y_{it}$, $P_{it}$, $OP_t$) of tourism demand should include as the determinants when using copula-ARDL and copula-ECM to generate multi-step forecasts. The calculation results are shown in Table 4. The significant dependence parameter $\theta$ of copula indicates that there is a positive correlation between the pairwise tourist flows for Group 1. This means that China, Japan, and South Korea hold a similarly positive impact towards Thailand, that is, the increase in the number of tourists from one country to Thailand will drive the increase of other two countries. Moreover, the demand of outbound tourists will be affected by some common factors for Group 1 countries, such as income, prices, natural disasters or economic crises and other emergencies. The results of Group 2 show that only the tourism flow between Malaysia and Singapore exists a positive impact. Laos and other two countries exist a negative influence on each other. As a result,
managers who formulate relevant tourism packaging policies can implement bundling sales in countries that choose to impact on each other positively.

Table 4. Estimated results for each pair of countries.

| Dependence parameter Θ         | Group 1  | Group 2  |
|-------------------------------|----------|----------|
| CN&JP                         | 0.359    | -0.330   |
| CN&KR                         | 0.700    | -0.383   |
| JP&KR                         | 0.645    |          |
| MA&SG                         | 0.824    |          |
| MA&LA                         | -0.330   |          |
| SG&LA                         | -0.383   |          |

Figure 1 shows the comparison between the forecasting out-of-sample flow and actual value of the six countries. Except Singapore, each country has shown good trends after adding the dependence structure of tourist flow. The performs in Figure 1 are consistent with results in Table 5. We evaluate the performance of ARDL and ECM copula models by using average absolute percentage error (MAPE) and root mean square error (RMSE) in the 4th, 12th, and 24th quarter-ahead forecasts to get Table 5. We know that MAPE values below 50% are acceptable, and MAPE values below 20% correspond to good results, so the MAPEs in the Table 5 can show that the forecasts of these models are acceptable and consistent with RMSEs.

Table 5. MAPE and RMSE of forecasting tourists from 6 countries.

|        | 4q MAPE | 4q RMSD | 12q MAPE | 12q RMSD | 24q MAPE | 24q RMSD |
|--------|---------|---------|----------|----------|----------|----------|
| Group 1|         |         |          |          |          |          |
| CN     | 0.015   | 0.320   | 0.013    | 0.258    | 0.013    | 0.256    |
| JP     | 0.013   | 0.202   | 0.051    | 0.251    | 0.097    | 0.239    |
| KR     | 0.013   | 0.179   | 0.013    | 0.177    | 0.079    | 0.188    |
| MAS    | 0.023   | 0.322   | 0.015    | 0.226    | 0.014    | 0.216    |
| Group 2|         |         |          |          |          |          |
| SG     | 0.409   | 5.147   | 1.108    | 4.702    | 1.847    | 3.994    |
| LA     | 0.025   | 0.353   | 0.018    | 0.269    | 0.080    | 0.210    |
5. Conclusion
This paper provides a comprehensive method to forecast the dependence structure of tourist flows between countries and the forecasting performs well. We know that the Copula is a multivariate probability distribution, where the marginal probability distribution of each variable is uniform. According to Sklar's theorem, any multivariate joint distribution can be written as a univariate marginal distribution function and a copula that describes the dependent structure between variables. So, it can accommodate different structures and relationships between flows. The pairwise correlation patterns of this copula-based framework can be extended to capture the correlation of three or more flows within the same region. But when there are more than two variables, the copula is less universal than two. In order to avoid this problem, parameters unconstrained optimization of copula is applied without repeated evaluation, which can improve the calculation efficiency, and obtain accurate forecasting.

Moreover, this study is also to prove that it is important to introduce the dependency structure into the tourism demand model. This result helps destination managers and travel agencies understand the impact of pairwise tourism flows on Thailand inbound tourist demand. In addition, it can also help the government to identify regional markets and implement tourism for package promotion in the same region. This provides a supplement for evaluating the role of tourism dependency structure in the international tourism demand model.

References
[1] Ontheera H and Naragain P 2019 Forecasting International Tourist Arrivals from Major Countries to Thailand Advances in Electronics Engineering 619 115-125.
[2] Puarattanaarunkorn O and Sriboonchitta S 2014 Modeling Dependence in Econometrics. Springer 251 383-398
[3] Zhu L and Zhang J 2017 Forecasting Chinese inbound tourism demand with Bernstein Copula function Tourism Tribune 32(11) 41-48.
[4] Song H, R Qiu, J Park. A 2019 review of research on tourism demand forecasting: Launching the Annals of Tourism Research Curated Collection on tourism demand forecasting Annals of Tourism Research 75 338–362.
[5] Song H and Li G 2008 Tourism demand modelling and forecasting–A review of recent research Tourism Management 29(2) 203–220.
[6] Lin V and Song H 2015 A review of Delphi forecasting research in tourism Current Issues in Tourism 18(12) 1099-1131.
[7] Li H, Hu M and Li G 2020 Forecasting tourism demand with multisource big data Annals of Tourism Research 83 102-912
[8] Calantone RJ, Di Benedetto A and Bojanic DC 1988 Multimethod forecasts for tourism analysis. Annals of Tourism Research 15(3) 387–406.
[9] Lim, C 1997 Review of international tourism demand models. Annals of Tourism Research 24(4) 835-849.
[10] Song H, Li G, Witt S F and Athanasopoulos G 2011 Forecasting tourist arrivals using time-varying parameter structural time series models. International Journal of Forecasting 27 855-869.
[11] Tang J, Sriboonchitta S,Ramos V and Wong W 2014 Modelling dependence between tourism demand and exchange rate using the copula-based GARCH model Current Issues in Tourism. 876-894
[12] Jorge V, Francisco L and Maria S 2015 Testing dependence between GDP and tourism's growth rates. Tourism Management 48 268-282
[13] Athanasopoulos G and Silva A 2012 Multivariate exponential smoothing for forecasting tourist arrivals Journal of Travel Research 51(5) 640–652.
[14] Zhu L, Lim C, Xie W and Wu Y 2018 Modelling tourist flow association for tourism demand forecasting. Current Issues in Tourism 21(8) 902–916.