Annals of Tourism Research 75 (2019) 363–378

Contents lists available at ScienceDirect

Annals of Tourism Research

journal homepage: www.elsevier.com/locate/annals

The combination of interval forecasts in tourism☆

Gang Li[a], Doris Chenguang Wu[b,*], Menglin Zhou[c], Anyu Liu[a]

[a] School of Hospitality and Tourism Management, University of Surrey, Guildford GU2 7XH, UK
[b] Business School, Sun Yat-sen University, Guangzhou, China
[c] Department of Statistics, University of British Columbia, Vancouver, Canada

ARTICLE INFO

Associate editor: Haiyan Song

Keywords:
Interval forecast
Combination forecasting
Econometric model
Winkler score
Tourism demand

ABSTRACT

Combination is an effective way to improve tourism forecasting accuracy. However, empirical evidence is limited to point forecasts. Given that interval forecasts can provide more comprehensive information, it is important to consider both point and interval forecasts for decision-making. Using Hong Kong tourism demand as an empirical case, this study is the first to examine if and how the combination can improve interval forecasting accuracy for tourism demand. Winkler scores are employed to measure interval forecasting performance. Empirical results show that combination improves the accuracy of tourism interval forecasting for different forecasting horizons. The findings provide government and industry practitioners with guidelines for producing accurate interval forecasts that benefit their policy-making for a wide array of applications in practice.

This article also launches the Annals of Tourism Research Curated Collection on Tourism Demand Forecast, a special selection of research in this field.

Introduction

Producing accurate tourism demand forecasts is of continuing interest to market practitioners. For example, hotels and airlines use demand forecasts to formulate their pricing and revenue management strategies; event organisers use them to allocate their resources more efficiently when organising festivals; governments use them to evaluate the feasibility of infrastructure investment. In recent years tourism demand forecasting has received even more attention from industry practitioners because of the continuously growing and more dynamic markets, decision makers' increasing realization of the importance of quantitative evidence for policy making, and the development of internet technology-based big data (Li & Wu, 2019). Due to the important role of tourism demand forecasting in facilitating business planning and policy formulation, a large number of empirical studies examine the accuracy of tourism demand forecasts using various forecasting techniques. The technique of forecast combination aims to combine the forecasts generated by a certain number of single models based on a certain weighting scheme. Empirical evidence indicates that forecast combination is an effective tool for improving forecast accuracy (Shen, Li, & Song, 2011; Song, Witt, Wong, & Wu, 2009; Wong, Song, Witt, & Wu, 2007).

According to a thorough review of the tourism forecasting literature (Wu, Song, & Shen, 2017), the majority of studies focus on point forecasting over interval forecasting: the former uses a single value to present the forecast, while the latter uses a range of values as the forecast outcome. In practice, point forecasts receive more attention, perhaps because they are easy to understand and their policy-making use is straightforward by providing a single value for a future situation. However, solely considering point

☆ The authors would like to acknowledge the financial support of the National Natural Science Foundation of China (Grant No. 71573289).
* Corresponding author.
E-mail addresses: g.li@surrey.ac.uk (G. Li), wucheng@mail.sysu.edu.cn (D.C. Wu), menglin.zhou@stat.ubc.ca (M. Zhou), anyu.liu@surrey.ac.uk (A. Liu).

https://doi.org/10.1016/j.annals.2019.01.010
Received 14 August 2018; Received in revised form 20 January 2019; Accepted 21 January 2019

forecasts is sometimes not enough for policy-making because they do not include any information on the variability associated with the forecasts. The use of both point forecasts and interval forecasts can provide more comprehensive and useful information for decision-making. For example, if a hotel has a point forecast of the rooms occupied on a future day, the hotel can predict and evaluate the occupancy rate and revenue/profit for that day; if the hotel also has interval forecasts, it knows the ranges that their occupancy rates and revenues may fall within under different confidence levels, and it can therefore formulate relevant pricing and marketing strategies. It is important for decision-makers to realise the importance of interval forecasts and to apply both point and interval forecasts as the foundation for policy-making. Accordingly, scholars should also explore effective ways to produce accurate interval forecasts. To the best of our knowledge, no study in the tourism literature has yet examined the performance of combining interval forecasts.

This study aims to fill the above gap by examining whether the combination forecasting technique can improve the accuracy of tourism demand interval forecasts in an empirical setting, considering Hong Kong's inbound tourism demand from its eight key source markets. Combined intervals are obtained by combining different density forecasts generated by eight individual time-series or econometric models. Two forecasting horizons of 1-step and 4-step ahead are examined separately. This study seeks to answer three research questions: how to combine interval forecasts? How can the accuracy of interval forecasts be evaluated effectively? Can combined interval forecasts improve forecasting accuracy over single forecasts?

The rest of the paper is structured as follows. Section 2 reviews the literature on tourism interval forecasting, tourism combination forecasting, interval forecast combination and forecasting accuracy measurement. Section 3 focuses on the methodology by introducing eight individual models, the interval combination technique, accuracy measures and data and variables for the empirical analysis. Section 4 presents the empirical results. Section 5 concludes the study and identifies future research directions.

Literature review

Tourism interval forecasting

Point estimation and point forecasting dominate the recent tourism forecasting literature (Wu et al., 2017). A point forecast uses a single value to forecast tourism demand for a certain point in the future. Most tourism forecasting studies use forecasting techniques such as time series or econometric models to produce point forecasts for different future time horizons. Recent examples include such as Athanasopoulos, Song, and Sun (2018) and Chen, Li, Wu, and Shen (2019). In contrast, interval forecasting provides a range instead of a single value to forecast at a given confidence level (or probability). As point forecasts do not provide information about the degree of variability or uncertainty of the forecast, interval forecasts are considered an effective supplement. Interval forecasts provide information not only on the central tendency of the forecast, but also on the future variation by constructing a range of values at a certain confidence level. The produced intervals can provide policymakers with more information and confidence. Furthermore, various intervals can be generated at different confidence levels set by forecasters in advance, enabling decision-makers to formulate policies or strategies based on different confidence levels.

Despite the advantages of interval forecasts, little attention has been paid to the application of interval forecasts to tourism demand, with some exceptions such as Kim, Song, and Wong (2010), Kim, Wong, Athanasopoulos, and Liu (2011) and Athanasopoulos, Hyndman, Song, and Wu (2011), among others. Kim et al. (2010) propose the bias-corrected bootstrap interval forecast of an autoregressive time series, and their empirical results in tourism demand show that this technique has desirable small-sample properties. Kim et al. (2011) compare tourism prediction intervals generated from alternative time series models, and they show that most models produce satisfactory intervals in terms of coverage values and width. Athanasopoulos et al. (2011) compare the coverage probabilities in tourism forecasting competition and demonstrate that the use of lower-frequency data tends to overestimate the coverage probabilities. Overall, interval forecasting remains an under-researched area in the tourism literature.

Tourism combination forecasting

Combination forecasting generates forecasts by combining the forecasts from a number of individual models using certain weighting methods. A large body of research indicates that the combination technique may improve the accuracy of point forecasting (see a review of Clemen, 1989). Some studies use combination forecasting technique in the tourism context, such as Shen, Li, and Song (2008); Shen et al. (2011), Song et al. (2009) and Wong et al. (2007). Empirical evidence in tourism forecasting suggests that no single model can generate the most accurate forecasts on all occasions (Song & Li, 2008; Wu et al., 2017). Combination forecasting avoids the risk of forecasting failure caused by relying on a single inappropriate forecasting model (Wong et al., 2007). Song et al. (2009) and Shen et al. (2011) demonstrate that the accuracy of combined forecasts is significantly higher than the average accuracy of individual forecasts. Although combination techniques have been applied to tourism forecasting, these applications are limited to combining point forecasts. Thus far, no study has explored how combined interval forecasts can be used in the tourism context.

Combination of interval forecasts

Combining interval forecasts provides not only a centre of intervals but also variability at certain confidence levels, and it is thus complex. It is straightforward to produce a combined interval forecast by combining the lower limits and upper limits respectively, yet this does not guarantee an interval with the correct probability (Timmermann, 2006). Therefore, we are unable to interpret combined intervals in practice. This problem has two solutions. One is to apply the quantile regression averaging (QRA) method (Liu,

Nowotarski, Hong, & Weron, 2017; Nowotarski & Weron, 2015), which provides a prediction interval by using quantile regression on point forecasts from individual methods to effectively construct a combined interval with the correct probability. The other is to derive the combined intervals from the combined density (Wallis, 2005). After obtaining the density function of each individual forecast, these density functions can be combined to generate forecasting intervals for any required probability; in contrast, the QRA method can only obtain an interval with a specific probability at a one-time combination. We therefore use the combined densities to produce the intervals in this study.

A density forecast of a random variable at some future time is an estimate of the probability distribution of this variable's possible future values; this estimate provides a complete description of the uncertainty associated with a prediction (Tay & Wallis, 2000). Researchers have developed various weighting schemes to combine density forecasts. For instance, Hendry and Clements (2004) advocate putting equal weights on all of the single interval forecasts involved. Granger and Jeon (2004) suggest a thick-modelling approach that eliminates the *m*% worst performing forecast and then takes a simple average of the remaining forecasts. Garratt, Lee, Pesaran, and Shin (2003) propose a Bayesian model averaging method that offers a means of weighting alternative models based on density forecasts according to their posterior probabilities. Hall and Mitchell (2007) obtain 'optimal' weights based on past forecast performance measured by minimizing the Kullback-Leibler information criterion (KLIC), which can measure the 'distance' between two densities. Garratt, Mitchell, Vahey, and Wakerly (2011) apply recursive weights to density forecasts, which use the logarithmic score to measure density performance. Billio, Casarin, Ravazzolo, and van Dijk (2013) propose a general distributional state space representation of predictive densities and combination schemes that can obtain time-varying weights with non-linear filtering.

A number of studies identify the advantage of density forecasting combination. Kascha and Ravazzolo (2012) point out that although combinations do not always outperform individual models, they are beneficial because they are more accurate overall and provide insurance against inappropriate model selection, which is the same as point forecasting combination. Since the combination of interval forecasts has not been examined in the field of tourism and hospitality and deserves attention (Wu et al., 2017), the current study examines whether tourism forecasting accuracy can be improved by combining interval forecasts.

Forecasting performance assessment

A number of measures can be used to assess forecasting performance in the case of point forecasting, most of which measure the distance between forecast values and real values. The most widely used measures include the mean absolute percentage error (MAPE), the root mean square error (RMSE), the root mean square percentage error (RMSPE) and the mean absolute error (MAE) (Wu et al., 2017).

An interval forecast produces a range instead of a single value, and its accuracy measurement thus cannot follow the same method. The coverage rate and the interval width are often used to measure the accuracy of interval forecasts (Athanasopoulos et al., 2011). The coverage rate is the percentage by which the real values fall in the prediction intervals. A coverage rate closer to the nominal coverage rate suggests better model performance. The interval width is the length of the prediction intervals generated at a certain confidence level. When the prediction intervals of two models produce the same coverage rates, the model with the narrower width is considered to have better forecasting performance (Kim et al., 2011).

An ideal prediction interval contains a coverage rate close to the nominal coverage rate and a narrow width. However, in practice, interval forecasts with superior coverage rates often have broader widths, which complicate the model choice. To consider both the coverage rate and the interval width for interval forecasting assessment, a comprehensive measurement, Winkler score (Winkler, 1972), is used, which penalises for observations outside the constructed interval and rewards for narrow widths. This study is the first to use the Winkler score to evaluate tourism interval forecasting performance.

Methodology

Individual forecasting models

Eight forecasting techniques that are widely used in the tourism demand forecasting literature are adopted in this study to generate individual forecasts: four time series models (naïve, exponential smoothing (ES), seasonal autoregressive integrate moving average (ARIMA) and structural time series (STS)) and four econometric approaches (autoregressive distributed lag (ADL), vector autoregressive (VAR), error correction (EC) and time-varying parameter (TVP)) models. The selection of individual models aims to cover a wide range of methods commonly applied in tourism forecasting with different merits. For example, the seasonal ARIMA model addresses seasonality, the ES and STS models in state space form decompose a time series into trend and seasonal components, the ADL model captures the long-term effect and dynamics of the demand system, the EC model focuses on the short-term effect, the VAR model uses systematic estimation, the TVP model estimates time-varying parameters, and the naïve model is commonly used as a benchmark in forecasting exercises (Li, Song, & Witt, 2005).

Naïve model

The naïve model states that future forecasts are simply equal to the recent available value. For yearly data, $\widehat{y}_t = y_{t-1}$, where y is the tourism demand, $\widehat{y}$ is its forecast and t refers to time. For data with seasonality, such as quarterly data, $\widehat{y}_t = y_{t-s}$, where s is 4. In this study, we use the naïve method with seasonality.

ES model

An exponential smoothing (ES) model consists of a trend, a seasonal component and additive or multiplicative errors. Hyndman, Koehler, Snyder, and Grose (2002) propose innovations state space models for exponential smoothing, which can be labelled as ETS (· ,· ,·). Particularly, three components need to be specified: the error type ("A" or "M"), the trend type ("N", "A" or "M"), and the season type ("N", "A" or "M") where "N" stands for none, "A" stands for additive and "M" represents multiplicative. An advantage of this ETS forecasting framework is that information criteria can be used for model selection. In our setting, models are selected by minimizing Bayesian information criterion (BIC).

Seasonal ARIMA model

The seasonal ARIMA approach is based on the standard Box-Jenkins method, and includes seasonal autoregressive and seasonal moving average structures. It is classified as a general-to-specific method in which all of the potential components are involved in the first-step model and all of the significant terms remain through a stepwise process based on the BIC. The seasonal ARIMA (p,d,q) $(P,D,Q)_s$ is given by

$$\phi(B)(1-B)^d\Phi(B^s)^D y_t = \theta(B)\Theta(B^s)\varepsilon_t, \tag{1}$$

where ε_t is the white noise with mean zero, B is the backshift operator, ϕ, θ, Φ and Θ are the polynomials of order p, q, P and Q, respectively, s denotes season, d is the difference operator and D is the seasonal difference operator. We use the 'auto.arima' function in the 'forecast' package in R to estimate the model.

STS model

The STS model proposed by Harvey (1989) consists of a stochastic trend, a seasonal term, an irregular component and a cyclical component. The STS model is specified as

$$y_t = \mu_t + \phi_t + \psi_t + \varepsilon_t, \tag{2}$$

where μ_t is the local linear trend component, which is assumed to follow a random walk, ϕ_t represents the seasonal component, which is defined to follow a stochastic specification (Vu & Turner, 2006), ψ_t is the cyclical component, which follows a trigonometric form (Harvey & Jaeger, 1993), and ε_t is the irregular component. In this study, we use an STS model with explanatory variables (Cortés-Jiménez & Blake, 2011), which is written as follows:

$$y_t = \mu_t + \phi_t + \psi_t + \sum_{i=1}^{k}\beta_i x_{i,t} + \varepsilon_t, \tag{3}$$

where $x_{i,\,t}$ is one of k explanatory variables. In this study, we construct the STS model according to Eq. (3) without the cyclical component (Kim et al., 2011).

VAR model

VAR model is a stochastic process used to capture the linear interdependencies among multiple time series. It generalises the univariate AR model by allowing more than one evolving variable. All of the variables in the VAR model join the model in the same way: each has an equation interpreting its evolution based on its own lagged value, the lagged values of other variables and an error term. The VAR model of order p is written as follows:

$$\boldsymbol{Y}_t = \boldsymbol{\Phi}_0 + \boldsymbol{\Phi}_1\begin{pmatrix} y_{t-1} \\ \boldsymbol{x}_{t-1}\end{pmatrix} + \ldots + \boldsymbol{\Phi}_p\boldsymbol{Y}_{t-p} + \boldsymbol{\varepsilon}_t, \tag{4}$$

where $\boldsymbol{Y}_t$ is the vector of endogenous variables, p is the lag order, Φ_i s are the constant matrices of the coefficients and $\boldsymbol{\varepsilon}_t \sim \mathrm{NID}(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$. In our setting, the lag order is 2 and the vector of endogenous variables includes tourist arrivals, own price and substitute price variables.

ADL model

The estimation of a dynamic econometric model proposed by Hendry (1986) is known as the general-to-specific approach (Shen et al., 2008; Wong et al., 2007). This method starts with a general ADL model:

$$y_t = \beta_0 + \sum_{j=1}^{k}\sum_{i=0}^{p_j}\beta_{j,i}x_{j,t-i} + \sum_{i=1}^{p}\phi_i y_{t-i} + \varepsilon_t, \tag{5}$$

where y_t is the tourism demand variable, x_j is the jth explanatory variable, p is the lag order of the dependent variable, p_j is the lag order of the jth regressor, $\beta_{j,i}$ and ϕ_i are coefficients and ε_t is white noise. After estimating the general ADL model, the most insignificant variable is removed from the equation and the model is re-estimated. This process is repeated until the variables remaining in

the model are all statistically significant and their estimates have correct signs in line with economic theory.

EC model

The EC model uses the Engle and Granger two-stage approach (Engle & Granger, 1987). The first step is to estimate a long-run cointegration regression model with the ordinary least squares method:

$$y_t = \beta_0 + \sum_{i=1}^{k} \beta_i x_{i,t} + u_t, \tag{6}$$

where u_t is the error term and β_i is the ith coefficient.

In the second step, the long-run cointegration relationship is transformed into an EC procedure with the term $\left(y_{t-1} - \beta_0 - \sum_{i=1}^{k} \beta_i x_{i,t-1}\right)$, and the EC model takes the following form:

$$\Delta y_t = \sum_{i=1}^{p} \phi_i \Delta y_{t-i} + \sum_{j=1}^{k} \sum_{i=0}^{q_j} \varphi_i \Delta x_{i,t-i} + \alpha\left(y_{t-1} - \beta_0 - \sum_{i=1}^{k} \beta_i x_{i-1}\right) + \varepsilon_t, \tag{7}$$

where ε_t is the error term and ϕ_i and β_is are coefficients.

TVP model

Different from the methods mentioned above, which assume that the coefficients of the variables are constant over time, the TVP model allows coefficients to change over time. In this study the autoregressive version of TVP model is used and expressed in the state space form

$$y_t = \beta_{0,t} + \sum_{i=1}^{p} \beta_{i,t} y_{t-i} + \sum_{j=p+1}^{p+k} \beta_{j,t} x_{j,t} + \varepsilon_t, \tag{8}$$

$$\beta_{i,t} = \beta_{i,t-1} + \omega_{i,t}, \tag{9}$$

where $p = 2$ in this study. The first equation is the measurement or system equation and the second is the transition or state equation. The Kalman filter algorithm (Kalman, 1960) is used to estimate this model.

Interval combination technique

In this study, the combined density is used to obtain the intervals. The advantage of this method is that once a density function is available, interval forecasts for any required probability can be obtained with the inverse function of the distribution function.

A density forecast of a random variable at some future time is an estimate of the probability distribution of the possible future values of that variable. Wan, Song, and Ko (2016) highlight the importance of density forecasts and the associated evaluation tool. In this study the forecast density functions for individual models are assumed to follow normal distribution, with point forecasts as the means. To estimate the variances, all the data sample is split into a training set, a validation set and a test set. The variances of the forecast density functions in the test set are estimated by the average squared difference between forecasts and actual values from the validation set.

Consider N density forecasts of variable y_t at time t, and denote these forecasts as $g_{i,t}$. Then, the linear combination of density forecasts is defined in the finite mixture:

$$f_t(y_t) = \sum_{i=1}^{N} w_i \cdot g_{i,t}(y_t), \tag{10}$$

where w_i are a set of non-negative weights that sum to 1. In these individual density forecasts, $m_{i,t}$ is the mean and $v_{i,t}$ is the variance. Further characteristics of combined density $f_t(y_t)$ can be expressed by the following:

$$E[f_t(y_t)] = m_t^* = \sum_{i=1}^{N} w_i \cdot m_{i,t}, \tag{11}$$

$$\mathrm{Var}[f_t(y_t)] = \sum_{i=1}^{N} w_i \cdot v_{i,t} + \sum_{i=1}^{N} w_i \cdot (m_{i,t} - m_t^*)^2. \tag{12}$$

Following Wallis (2005), we apply equal weights to the above functions:

$$f_t(y_t) = \sum_{i=1}^{N} \frac{1}{N} \cdot g_{i,t}(y_t), \tag{13}$$

$$E[f_t(y_t)] = m_t^* = \frac{1}{N}\sum_{i=1}^{N} m_{i,t}, \tag{14}$$

$$\mathrm{Var}[f_t(y_t)] = \frac{1}{N}\sum_{i=1}^{N} v_{i,t} + \frac{1}{N}\sum_{i=1}^{N} (m_{i,t} - m_t^*)^2. \tag{15}$$

With the combined distribution function $f_t(y_t)$, we denote the quantile function as $Q_t(\theta) = f_t^{-1}(\theta)$. A central interval with confidence level p is defined as $[Q_t(q/2), Q_t(1 - q/2)]$, where $q = 1 - p$. Given the difficulty of calculating the analytic integral of a normal density function directly, we use the Taylor expansion to help determine the approximate value.

Forecasting accuracy measurement

In point forecasting, MAPE is used to measure the forecasting performance with the following formula:

$$MAPE = \frac{1}{m}\sum_{t=1}^{m} \frac{|y_t - \hat{y}_t|}{y_t} \times 100\%. \tag{16}$$

In interval forecasting, the width, coverage rate and Winkler score are used to measure forecasting performance. Suppose that we have m prediction intervals $\{[\widehat{L}_t, \widehat{U}_t], t = 1,..., m\}$ for the m true future values of $\{y_t, t = 1, \dots, m\}$, where L_t is the lower limit and U_t is the upper limit. The width of an interval forecast (W_t) is defined as follows:

$$W_t = \widehat{U}_t - \widehat{L}_t, \tag{17}$$

while the coverage rate is written as

$$C = \frac{\#\{\hat{L}_t < y_t < \hat{U}_t\}}{m}, \tag{18}$$

where # is the frequency at which the condition inside the bracket is satisfied. The prediction interval whose true coverage rate is close to the nominal coverage rate is preferred. When two prediction intervals have similar coverage rates, the one with a narrower width is favoured.

In practice, the coverage rate and width may lead to controversial evaluation results. Therefore, the Winkler score, a comprehensive measure, is used to jointly assess the interval width and the coverage rate. For a central $p \times 100\%$ prediction interval of value y_t, the Winkler score is defined as follows:

$$\text{Winkler} = \begin{cases} W_t & L_t \le y_t \le U_t \\ W_t + 2(L_t - y_t)/(1 - p) & L_t > y_t \\ W_t + 2(y_t - U_t)/(1 - p) & U_t < y_t. \end{cases} \tag{19}$$

The Winkler score gives a penalty if y_t is outside the prediction interval (Hong & Fan, 2016), and a lower score indicates a better prediction interval. This study is the first to introduce the Winkler score to tourism interval forecasting performance evaluation.

Particularly, when measuring the forecasting accuracy, MAPEs of point forecasts are calculated based on the original scale of the tourism demand. But for interval forecasts, since forecasts of the tourism demand in natural logarithm is assumed normally distributed, and interval combination is based on these normal distributions, we measure interval forecasting accuracy based on the natural logarithm form of the tourism demand instead of the original scales.

Data description and modelling process

This study focuses on Hong Kong's inbound tourism demand from its eight key source markets: mainland China, Taiwan, South Korea, Japan, Macao, the Philippines, Singapore and the US. Fig. 1 shows the time series plots of natural log-transformed tourism arrivals from eight source markets. In line with prior studies such as Song, Wong, and Chon (2003), own price, substitute price and income are used as determinants of tourism demand in this study. The own price p_{it} and substitute price p_{st} are defined as

$$p_{it} = \frac{\mathrm{CPI}_{HK}/\mathrm{EX}_{HK}}{\mathrm{CPI}_{origin}/\mathrm{EX}_{origin}}, \tag{20}$$

$$p_{st} = \sum_{j=1}^{4} \frac{\mathrm{CPI}_j}{\mathrm{EX}_j} w_j, \tag{21}$$

where CPI and EX denote the customer price index and the exchange rate, respectively, and j denotes the jth substitute destination for Hong Kong. Four substitute destinations are selected: South Korea, Japan, Macao and Singapore. w_j is the share of tourist arrivals in the jth substitute destination for the total tourist arrivals in these four countries. In each of the models using South Korea, Japan, Macao and Singapore as source markets, the same country is excluded from the calculation of the substitute price variable for that model. The income variable is measured by the real GDP index in the constant prices of these eight source markets. Seasonal dummies

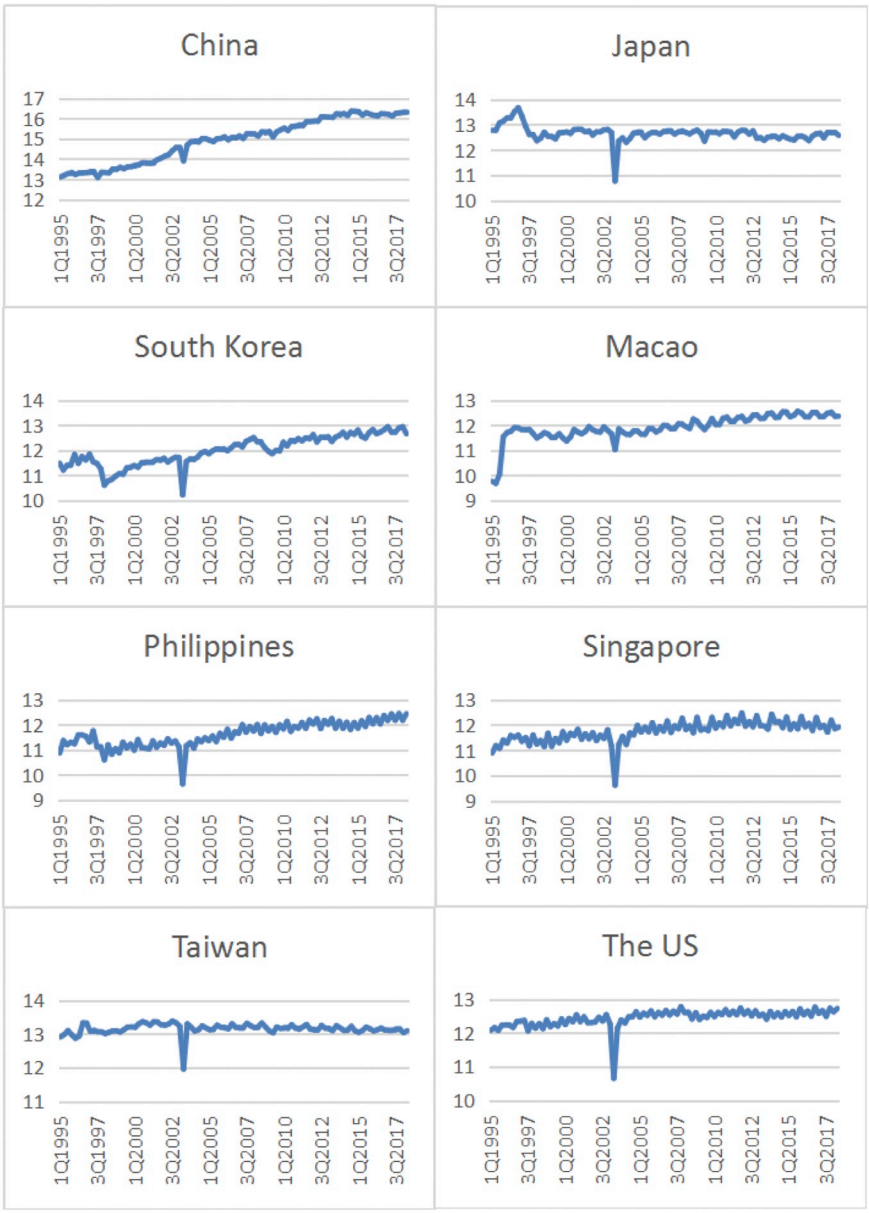

Fig. 1. Time series plots for log-transformed tourist arrivals.

are added to capture the seasonal effects, and one-off event dummies are included to reflect the effects of SARS in 2003 and the financial crisis in 2009. The sample covers the 1995Q1-2018Q2 period. All of the variables except the dummies are transferred to a natural logarithm before the model estimation, and the rolling window method is used to produce both point and interval forecasts.

The individual model forecasting process follows the following steps. First, we use the training dataset (1995Q1-2008Q2) for model estimation. We adopt an increasing rolling window to our training set to produce one-step-ahead forecasts for 2008Q3-2018Q2 (Wong et al., 2007). Second, the validation dataset (2008Q3-2013Q2) aims to compute variances for density forecasts. For example, the residuals for the period of 2008Q3-2013Q2 are used to estimate variances for the density forecasts of 2013Q3. A rolling window is adopted to our validation set until the variances for 2018Q2 are obtained. Finally, the test dataset (2013Q3-2018Q2) is used to generate prediction intervals of individual models, and both 80% and 70% confidence levels are used for each model and each source market. Following the same procedures, we further produce prediction intervals for two- to four-step ahead forecasting horizons. Due to space constraints, only one- and four-step ahead results are presented in the following section.

Once the density functions of all individual models are forecast, all possible combinations for these density functions are conducted for each source market from a two-model combination to an eight-model combination. Based on these combinations, Taylor's expansion is used to calculate the approximation of prediction on intervals for each combining model.

Table 1
MAPE values for single and combined forecasts (%).

	China	Japan	Korea	Macao	Philippines	Singapore	Taiwan	USA	Average
One-step-ahead									
Single models:									
NAIVE	8.78	7.78	9.05	3.42	8.95	8.75	3.17	3.55	6.68
ES	6.14	8.16	11.34	5.62	11.82	7.84	5.22	8.92	8.13
ARIMA	4.27	5.84	8.41	4.29	7.06	8.41	2.24	2.91	5.43
STS	4.17	4.72	8.09	4.13	7.38	8.40	2.62	2.84	5.30
VAR	9.03	8.08	13.34	8.92	16.75	24.77	4.56	10.05	11.94
ADL	7.61	8.42	7.72	11.71	12.02	12.69	9.93	8.14	9.78
EC	5.12	4.61	7.83	4.96	11.07	6.94	3.81	8.51	6.61
TVP	3.66	10.02	11.49	3.44	12.34	12.12	3.76	6.82	7.96
Average	**6.10**	**7.20**	**9.66**	**5.81**	**10.92**	**11.24**	**4.42**	**6.47**	**7.73**
Average MAPEs of different combined forecasts:									
Two-model	5.22	5.77	9.05	5.10	10.04	9.83	3.69	6.01	6.84
Three-model	4.83	5.21	8.82	4.77	9.78	9.35	3.38	5.86	6.50
Four-model	4.60	4.91	8.68	4.61	9.65	9.10	3.20	5.79	6.32
Five-model	4.47	4.71	8.59	4.54	9.59	8.98	3.09	5.75	6.22
Six-model	4.37	4.58	8.53	4.49	9.56	8.90	3.01	5.73	6.15
Seven-model	4.30	4.50	8.48	4.43	9.56	8.84	2.94	5.72	6.10
Eight-model	4.24	4.45	8.45	4.38	9.56	8.81	2.91	5.71	6.06
Four-step-ahead									
Single models:									
NAIVE	7.96	6.89	9.31	2.93	9.20	8.61	3.36	3.34	6.45
ES	17.13	7.76	10.59	4.24	12.53	7.37	5.25	7.22	9.01
ARIMA	10.60	12.05	13.09	17.82	11.93	16.41	3.74	2.89	11.07
STS	10.87	9.48	11.06	15.02	12.17	10.91	4.06	2.60	9.52
VAR	17.33	13.11	16.99	5.44	15.69	25.38	4.41	8.26	13.33
ADL	9.08	10.55	8.16	11.52	14.23	14.87	10.52	6.75	10.71
EC	16.05	6.46	12.53	4.20	17.05	8.04	4.17	11.56	10.00
TVP	6.75	10.35	17.48	3.53	16.05	22.97	4.15	6.06	10.92
Average	**11.97**	**9.58**	**12.40**	**8.08**	**13.61**	**14.32**	**4.96**	**6.09**	**10.13**
Average MAPEs of different combined forecasts:									
Two-model	10.73	8.06	11.98	6.35	13.17	12.77	4.18	5.09	9.04
Three-model	10.21	7.52	11.83	5.40	13.04	12.21	3.88	4.65	8.59
Four-model	9.93	7.26	11.77	4.80	12.98	11.87	3.71	4.40	8.34
Five-model	9.76	7.10	11.74	4.40	12.94	11.65	3.61	4.25	8.18
Six-model	9.65	6.99	11.71	4.09	12.91	11.51	3.55	4.15	8.07
Seven-model	9.55	6.91	11.68	3.85	12.90	11.38	3.50	4.07	7.98
Eight-model	9.48	6.87	11.64	3.67	12.90	11.28	3.47	3.98	7.91

Finally, the forecasting performance is evaluated. The performance of point forecasts is examined using MAPE and the interval forecasting performance is examined using the interval width, coverage rate and Winkler score.

Empirical results

Point forecast combination

First, the combination based on point forecasts is conducted, and its performance is examined using MAPE values. The results are shown in Table 1, where the point forecasting performance for the eight individual models and the combined forecasts for the eight countries are reported. It is observed that the performances of these individual models vary across the eight source markets for both one-step ahead and four-step ahead forecasting horizons. For example, the TVP model performs best for the China case, while the EC model performs best for the Japan case.

Regarding point forecasting combination, to examine whether combination can improve the forecasting performance over single models, the average MAPEs are compared between single model forecasts and combined forecasts, as shown in Table 1. It is observed that for all eight source markets, the average MAPEs of the combined forecasts are lower than the average MAPEs of the single models for all of the combination cases, which indicates that combination is an effective way to improve point forecasting accuracy on average. Using the China case as an example, in the one-step ahead case, the average MAPEs for all of the combination forecasts range from 4.24% to 5.22%; all of these values are lower than the average MAPE of the single models of 6.10%. Thus, combination can improve point forecasting accuracy on average. We further calculate and compare the median MAPEs between single forecasts and combined forecasts and obtain the same conclusion. This further verifies the effectiveness of combination in improving point forecasting performance. Due to space constraints the results for median MAPEs are omitted.

Table 2
Coverage rates for single and combined interval forecasts (%).

	80% confidence level		70% confidence level	
	One-step-ahead	Four-step-ahead	One-step-ahead	Four-step-ahead
Single models:				
NAIVE	82.50	82.35	75.63	75.74
ES	85.00	77.94	66.25	64.71
ARIMA	83.13	55.88	73.75	47.79
STS	86.88	69.85	73.75	58.09
VAR	74.38	66.91	59.38	59.56
ADL	73.13	63.24	58.13	52.21
EC	80.63	65.44	73.13	55.15
TVP	71.88	60.29	64.38	55.15
Average	**79.69**	**67.74**	**68.05**	**58.55**
Average coverage rates of different combined forecasts:				
Two-model	82.81	72.53	75.83	66.57
Three-model	83.18	73.56	76.24	68.15
Four-model	82.55	73.63	76.54	68.69
Five-model	82.32	73.96	76.94	68.83
Six-model	82.01	73.90	77.19	68.80
Seven-model	81.41	73.35	77.19	68.75
Eight-model	81.88	78.13	78.13	67.65

Coverage rates and interval widths

Regarding the performance of interval forecasts, Tables 2 and 3 exhibit the averages of the coverage rates and interval widths across the eight source markets for the eight individual models at 80% and 70% confidence levels, respectively. It is observed that for one-step-ahead interval forecasts, the coverage rates are quite close to the nominal ones (with an average of 79.69% at the 80% confidence level and 68.05% at the 70% confidence level). Regarding the four-step-ahead interval forecasts, the coverage rates are generally quite below the nominal ones, with an average of 67.74% at the 80% confidence level, and 58.55% at the 70% confidence level. It is also observed that the coverage rates of combined interval forecasts are higher than the average ones of single models for both confidence levels and both forecasting horizons.

Table 3 reports the interval widths for forecasts of single models and combination forecasts. Although ideally, an interval forecast with the coverage rate closest to the nominal one and with the narrowest width is preferred, the trade-off between coverage rates and interval widths is often observed. For example, for the last columns in Tables 2 and 3, it is shown that combined intervals have superior coverage rates over the average of single models, but generally have wider interval widths. Thus, using both measures to evaluate the performance of interval forecasts may lead to controversial conclusions. To overcome this limitation, the Winkler score is used which takes account of the merits of both coverage rates and interval widths.

Table 3
Interval widths for single and combined interval forecasts.

	80% Confidence level		70% Confidence level	
	One-step-ahead	Four-step-ahead	One-step-ahead	Four-step-ahead
Single models:				
NAIVE	0.247	0.233	0.195	0.190
ES	0.272	0.267	0.217	0.218
ARIMA	0.195	0.232	0.154	0.189
STS	0.196	0.245	0.155	0.200
VAR	0.330	0.334	0.262	0.273
ADL	0.262	0.263	0.207	0.217
EC	0.219	0.247	0.174	0.202
TVP	0.221	0.233	0.176	0.191
Average	**0.243**	**0.257**	**0.193**	**0.210**
Average widths of different combined forecasts:				
Two-model	0.241	0.258	0.203	0.223
Three-model	0.238	0.253	0.202	0.222
Four-model	0.234	0.248	0.201	0.220
Five-model	0.232	0.243	0.200	0.217
Six-model	0.230	0.239	0.199	0.214
Seven-model	0.228	0.235	0.198	0.211
Eight-model	0.227	0.232	0.197	0.209

Table 4
Winkler scores for single and combined interval forecasts (one-step-ahead forecasts).

	China	Japan	Korea	Macao	Philippines	Singapore	Taiwan	US	Average	China	Japan	Korea	Macao	Philippines	Singapore	Taiwan	US	Average
	80% confidence level									70% confidence level								
Single models:																		
NAIVE	0.402	0.395	0.463	0.169	0.368	0.383	0.151	0.164	**0.312**	0.397	0.338	0.384	0.139	0.349	0.325	0.130	0.140	**0.275**
ES	0.279	0.297	0.375	0.290	0.485	0.363	0.199	0.333	**0.328**	0.237	0.268	0.360	0.262	0.437	0.318	0.175	0.296	**0.294**
ARIMA	0.196	0.293	0.429	0.240	0.302	0.297	0.127	0.128	**0.252**	0.159	0.254	0.359	0.211	0.264	0.287	0.111	0.109	**0.219**
STS	0.195	0.266	0.383	0.223	0.358	0.310	0.126	0.127	**0.249**	0.168	0.227	0.324	0.185	0.304	0.284	0.109	0.108	**0.214**
VAR	0.346	0.368	0.685	0.326	0.789	0.816	0.193	0.381	**0.488**	0.342	0.327	0.595	0.318	0.693	0.742	0.175	0.355	**0.443**
ADL	0.354	0.348	0.395	0.395	0.422	0.486	0.499	0.316	**0.402**	0.299	0.312	0.334	0.381	0.409	0.448	0.408	0.292	**0.360**
EC	0.211	0.213	0.391	0.221	0.439	0.275	0.191	0.404	**0.293**	0.192	0.177	0.338	0.192	0.411	0.245	0.165	0.348	**0.258**
TVP	0.195	0.410	0.689	0.196	0.653	0.593	0.172	0.284	**0.399**	0.163	0.353	0.560	0.171	0.572	0.495	0.152	0.250	**0.340**
Average	**0.272**	**0.324**	**0.476**	**0.258**	**0.477**	**0.440**	**0.207**	**0.267**	0.340	**0.245**	**0.282**	**0.407**	**0.232**	**0.430**	**0.393**	**0.178**	**0.237**	0.300
Percentage of two-model combination forecasts that outperform single-model forecasts (%), $n = 28$																		
Best	28.6	14.3	57.1	10.7	25.0	17.9	32.1	25.0	**26.3**	14.3	14.3	57.1	14.3	28.6	25.0	32.1	28.6	**26.8**
In-between	64.3	85.7	42.9	89.3	75.0	82.1	67.9	75.0	**72.8**	85.7	85.7	42.9	85.7	71.4	75.0	67.9	71.4	**73.2**
Worst	7.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.9**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
Percentage of three-model combination forecasts that outperform single-model forecasts (%), $n = 70$																		
Best	14.3	1.8	44.6	8.9	16.1	16.1	14.3	12.5	**16.1**	5.4	7.1	42.9	5.4	14.3	19.6	8.9	12.5	**14.5**
In-between	85.7	98.2	55.4	91.1	83.9	83.9	85.7	76.8	**82.6**	94.6	92.9	57.1	94.6	85.7	80.4	91.1	87.5	**85.5**
Worst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.7	**1.3**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
Percentage of four-model combination forecasts that outperform single-model forecasts (%), $n = 56$																		
Best	4.3	0.0	44.3	2.9	10.0	11.4	4.3	5.7	**10.4**	0.0	1.4	40.0	1.4	4.3	15.7	1.4	5.7	**8.8**
In-between	95.7	100.0	55.7	97.1	90.0	88.6	95.7	84.3	**88.4**	100.0	98.6	60.0	98.6	95.7	84.3	98.6	94.3	**91.3**
Worst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	**1.3**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
Percentage of five-model combination forecasts that outperform single-model forecasts (%), $n = 70$																		
Best	1.8	0.0	32.1	3.6	3.6	12.5	1.8	1.8	**7.1**	0.0	0.0	30.4	0.0	0.0	5.4	0.0	1.8	**4.7**
In-between	98.2	100.0	67.9	96.4	96.4	87.5	98.2	91.1	**92.0**	100.0	100.0	69.6	100.0	100.0	94.6	100.0	98.2	**95.3**
Worst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.1	**0.9**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
Percentage of six-model combination forecasts that outperform single-model forecasts (%), $n = 28$																		
Best	0.0	0.0	25.0	0.0	0.0	3.6	0.0	0.0	**3.6**	0.0	0.0	21.4	0.0	0.0	3.6	0.0	0.0	**3.1**
In-between	100.0	100.0	75.0	100.0	100.0	96.4	100.0	96.4	**96.0**	100.0	100.0	78.6	100.0	100.0	96.4	100.0	100.0	**96.9**
Worst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.6	**0.4**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
Percentage of seven-model combination forecasts that outperform single-model forecasts (%), $n = 8$																		
Best	0.0	0.0	12.5	0.0	0.0	0.0	0.0	0.0	**1.6**	0.0	0.0	12.5	0.0	0.0	0.0	0.0	0.0	**1.6**
In-between	100.0	100.0	87.5	100.0	100.0	100.0	100.0	100.0	**98.4**	100.0	100.0	87.5	100.0	100.0	100.0	100.0	100.0	**98.4**
Worst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
Percentage of eight-model combination forecasts that outperform single-model forecasts (%), $n = 1$																		
Best	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
In-between	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	**100.0**
Worst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
Highest Winkler scores of single and combined forecasts:																		
Single model	0.402	0.410	0.689	0.395	0.789	0.816	0.499	0.404	**0.550**	0.397	0.353	0.595	0.381	0.693	0.742	0.408	0.355	**0.490**
Two-model	0.354	0.392	0.644	0.359	0.686	0.724	0.353	0.367	**0.485**	0.306	0.339	0.547	0.336	0.596	0.592	0.277	0.321	**0.414**
Three-model	0.333	0.375	0.529	0.335	0.568	0.610	0.306	0.433*	**0.436**	0.295	0.317	0.454	0.294	0.519	0.527	0.238	0.319	**0.370**

(*continued on next page*)

Table 4 (*continued*)

	China	Japan	Korea	Macao	Philippines	Singapore	Taiwan	US	Average	China	Japan	Korea	Macao	Philippines	Singapore	Taiwan	US	Average
	80% confidence level									70% confidence level								
Four-model	0.292	0.359	0.478	0.302	0.511	0.526	0.271	0.466*	**0.400**	0.256	0.303	0.413	0.256	0.476	0.458	0.215	0.345	**0.340**
Five-model	0.264	0.345	0.451	0.256	0.479	0.460	0.217	0.458*	**0.366**	0.234	0.294	0.385	0.220	0.424	0.401	0.182	0.340	**0.310**
Six-model	0.251	0.335	0.438	0.235	0.431	0.411	0.182	0.419*	**0.338**	0.223	0.285	0.372	0.205	0.395	0.363	0.157	0.318	**0.290**
Seven-model	0.239	0.328	0.420	0.217	0.394	0.367	0.165	0.339	**0.309**	0.214	0.276	0.359	0.193	0.365	0.334	0.145	0.263	**0.269**
Eight-model	0.222	0.311	0.395	0.204	0.374	0.339	0.151	0.266	**0.283**	0.202	0.263	0.342	0.177	0.341	0.310	0.136	0.217	**0.248**
Average Winkler scores of single and combined forecasts:																		
Single model	0.272	0.324	0.476	0.258	0.477	0.440	0.207	0.267	**0.340**	0.245	0.282	0.407	0.232	0.430	0.393	0.178	0.237	**0.300**
Two-model	0.246	0.312	0.413	0.238	0.427	0.401	0.197	0.260	**0.312**	0.216	0.268	0.360	0.207	0.384	0.353	0.165	0.223	**0.272**
Three-model	0.237	0.312	0.399	0.233	0.407	0.376	0.183	0.270*	**0.302**	0.210	0.266	0.349	0.201	0.368	0.335	0.156	0.225	**0.264**
Four-model	0.232	0.312	0.394	0.225	0.397	0.361	0.171	0.275*	**0.296**	0.206	0.265	0.344	0.195	0.359	0.326	0.149	0.226	**0.259**
Five-model	0.229	0.311	0.393	0.219	0.391	0.352	0.163	0.276*	**0.292**	0.204	0.264	0.343	0.190	0.353	0.320	0.143	0.225	**0.255**
Six-model	0.227	0.311	0.393	0.214	0.384	0.346	0.157	0.274*	**0.288**	0.203	0.264	0.342	0.186	0.348	0.315	0.139	0.224	**0.253**
Seven-model	0.224	0.311	0.395	0.207	0.379	0.341	0.154	0.270*	**0.285**	0.203	0.264	0.342	0.182	0.345	0.313	0.137	0.221	**0.251**
Eight-model	0.222	0.311	0.395	0.204	0.374	0.339	0.151	0.266	**0.283**	0.202	0.263	0.342	0.177	0.341	0.310	0.136	0.217	**0.248**

Note: 'Best' refers to the combined forecasts that outperform the best individual forecasts among the component models in a combination.
'In-between' indicates the combined forecasts that are inferior to the best individual component forecasts and outperform the worst individual component forecasts.
"Worst" refers to the combined forecasts that cannot outperform the worst individual component forecasts.
*Refers to the situations that the Winkler scores for combined forecasts are higher than the average of individual component forecasts.

Table 5
Winkler scores for single and combined interval forecasts (four-step-ahead forecasts).

	China	Japan	Korea	Macao	Philippines	Singapore	Taiwan	US	Average	China	Japan	Korea	Macao	Philippines	Singapore	Taiwan	US	Average
	80% confidence level									70% confidence level								
Single models:																		
NAIVE	0.370	0.364	0.424	0.152	0.365	0.400	0.140	0.152	**0.296**	0.341	0.305	0.354	0.125	0.350	0.331	0.125	0.132	**0.258**
ES	0.689	0.352	0.419	0.185	0.507	0.321	0.208	0.307	**0.373**	0.577	0.308	0.410	0.170	0.474	0.280	0.179	0.272	**0.334**
ARIMA	0.489	0.502	0.789	0.795	0.485	0.695	0.168	0.129	**0.507**	0.409	0.445	0.635	0.659	0.450	0.600	0.135	0.115	**0.431**
STS	0.524	0.470	0.501	0.660	0.525	0.390	0.194	0.149	**0.426**	0.440	0.386	0.435	0.564	0.464	0.351	0.159	0.123	**0.365**
VAR	0.726	0.458	0.926	0.264	0.717	0.888	0.197	0.332	**0.563**	0.651	0.424	0.785	0.236	0.655	0.792	0.168	0.298	**0.501**
ADL	0.532	0.398	0.411	0.422	0.546	0.584	0.524	0.275	**0.462**	0.453	0.362	0.351	0.400	0.512	0.532	0.413	0.255	**0.410**
EC	0.578	0.306	0.628	0.163	0.726	0.448	0.170	0.257	**0.409**	0.556	0.269	0.554	0.142	0.667	0.382	0.148	0.197	**0.364**
TVP	0.388	0.421	1.139	0.178	0.935	1.046	0.193	0.287	**0.573**	0.325	0.370	0.893	0.146	0.802	0.866	0.157	0.246	**0.476**
Average	**0.537**	**0.409**	**0.655**	**0.352**	**0.601**	**0.597**	**0.224**	**0.236**	0.451	**0.469**	**0.359**	**0.552**	**0.305**	**0.547**	**0.517**	**0.186**	**0.205**	0.392
Percentage of two-model combination forecasts that outperform single-model forecasts (%), $n = 28$																		
Best	39.3	17.9	32.1	28.6	32.1	21.4	57.1	17.9	**30.8**	32.1	21.4	21.4	17.9	25.0	25.0	39.3	14.3	**24.6**
In-between	60.7	78.6	67.9	71.4	67.9	78.6	42.9	67.9	**67.0**	67.9	78.6	78.6	82.1	75.0	75.0	60.7	71.4	**73.7**
Worst	0.0	3.6	0.0	0.0	0.0	0.0	0.0	14.3	**2.2**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.3	**1.8**
Percentage of three-model combination forecasts that outperform single-model forecasts (%), $n = 70$																		
Best	32.1	19.6	25.0	12.5	19.6	19.6	37.5	10.7	**22.1**	23.2	14.3	16.1	8.9	14.3	12.5	28.6	10.7	**16.1**
In-between	67.9	80.4	75.0	87.5	80.4	80.4	62.5	89.3	**77.9**	76.8	85.7	83.9	91.1	85.7	87.5	71.4	89.3	**83.9**
Worst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
Percentage of four-model combination forecasts that outperform single-model forecasts (%), $n = 56$																		
Best	18.6	11.4	14.3	1.4	15.7	14.3	25.7	2.9	**13.0**	12.9	10.0	7.1	2.9	5.7	7.1	17.1	2.9	**8.2**
In-between	81.4	88.6	85.7	98.6	84.3	85.7	74.3	97.1	**87.0**	87.1	90.0	92.9	97.1	94.3	92.9	82.9	97.1	**91.8**
Worst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
Percentage of five-model combination forecasts that outperform single-model forecasts (%), $n = 70$																		
Best	12.5	3.6	8.9	0.0	14.3	7.1	16.1	0.0	**7.8**	5.4	5.4	3.6	0.0	1.8	3.6	5.4	0.0	**3.1**
In-between	87.5	96.4	91.1	100.0	85.7	92.9	83.9	100.0	**92.2**	94.6	94.6	96.4	100.0	98.2	96.4	94.6	100.0	**96.9**
Worst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
Percentage of six-model combination forecasts that outperform single-model forecasts (%), $n = 28$																		
Best	3.6	0.0	3.6	0.0	7.1	3.6	7.1	0.0	**3.1**	0.0	3.6	0.0	0.0	0.0	0.0	0.0	0.0	**0.4**
In-between	96.4	100.0	96.4	100.0	92.9	96.4	92.9	100.0	**96.9**	100.0	96.4	100.0	100.0	100.0	100.0	100.0	100.0	**99.6**
Worst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
Percentage of seven-model combination forecasts that outperform single-model forecasts (%), $n = 8$																		
Best	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
In-between	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	**100.0**	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	**100.0**
Worst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
Percentage of eight-model combination forecasts that outperform single-model forecasts (%), $n = 1$																		
Best	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
In-between	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	**100.0**
Worst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
Highest Winkler scores of single and combined forecasts:																		
Single model	0.726	0.502	1.139	0.795	0.935	1.046	0.524	0.332	**0.750**	0.651	0.445	0.893	0.659	0.802	0.866	0.413	0.298	**0.628**
Two-model	0.689	0.504*	0.923	0.731	0.790	0.835	0.394	0.362*	**0.654**	0.610	0.422	0.785	0.526	0.729	0.734	0.295	0.314*	**0.552**
Three-model	0.648	0.473	0.866	0.616	0.679	0.780	0.337	0.297	**0.587**	0.564	0.405	0.727	0.450	0.637	0.659	0.258	0.257	**0.495**

(*continued on next page*)

Table 5 (*continued*)

	China	Japan	Korea	Macao	Philippines	Singapore	Taiwan	US	Average	China	Japan	Korea	Macao	Philippines	Singapore	Taiwan	US	Average
	80% confidence level									70% confidence level								
Four-model	0.654	0.447	0.821	0.481	0.622	0.713	0.291	0.315	**0.543**	0.535	0.393	0.672	0.352	0.581	0.605	0.228	0.254	**0.453**
Five-model	0.564	0.431	0.773	0.348	0.587	0.613	0.230	0.292	**0.480**	0.475	0.367	0.635	0.271	0.544	0.533	0.189	0.231	**0.406**
Six-model	0.504	0.415	0.648	0.282	0.549	0.515	0.196	0.260	**0.421**	0.441	0.348	0.557	0.229	0.513	0.461	0.168	0.208	**0.366**
Seven-model	0.455	0.400	0.575	0.247	0.514	0.440	0.182	0.213	**0.378**	0.402	0.333	0.496	0.204	0.487	0.405	0.159	0.176	**0.333**
Eight-model	0.420	0.361	0.520	0.205	0.480	0.400	0.171	0.187	**0.343**	0.377	0.310	0.458	0.171	0.456	0.357	0.153	0.162	**0.306**
Average Winkler scores of single and combined forecasts:																		
Single model	0.537	0.409	0.655	0.352	0.601	0.597	0.224	0.236	**0.451**	0.469	0.359	0.552	0.305	0.547	0.517	0.186	0.205	**0.392**
Two-model	0.482	0.393	0.553	0.317	0.528	0.504	0.205	0.229	**0.401**	0.419	0.339	0.488	0.256	0.496	0.439	0.174	0.199	**0.351**
Three-model	0.461	0.385	0.528	0.277	0.506	0.466	0.191	0.215	**0.379**	0.401	0.330	0.472	0.227	0.477	0.412	0.165	0.186	**0.334**
Four-model	0.446	0.379	0.521	0.249	0.494	0.444	0.181	0.207	**0.365**	0.392	0.324	0.467	0.208	0.468	0.395	0.159	0.179	**0.324**
Five-model	0.436	0.374	0.518	0.230	0.487	0.428	0.175	0.199	**0.356**	0.386	0.320	0.465	0.194	0.463	0.383	0.155	0.174	**0.317**
Six-model	0.428	0.370	0.518	0.219	0.483	0.415	0.172	0.194	**0.350**	0.382	0.316	0.463	0.185	0.460	0.373	0.153	0.169	**0.313**
Seven-model	0.423	0.366	0.519	0.211	0.480	0.405	0.171	0.192	**0.346**	0.378	0.314	0.461	0.178	0.457	0.364	0.153	0.167	**0.309**
Eight-model	0.420	0.361	0.520	0.205	0.480	0.400	0.171	0.187	**0.343**	0.377	0.310	0.458	0.171	0.456	0.357	0.153	0.162	**0.306**

Note: 'Best' refers to the combined forecasts that outperform the best individual forecasts among the component models in a combination.
'In-between' indicates the combined forecasts that are inferior to the best individual component forecasts and outperform the worst individual component forecasts.
"Worst" refers to the combined forecasts that cannot outperform the worst individual component forecasts.
*Refers to the situations that the Winkler scores for combined forecasts are higher than the average of individual component forecasts.

Winkler scores for interval forecast combination

Lower Winkler scores mean better performance of interval forecasts. As Tables 4 and 5 illustrate, when single models are concerned, the single models with the best performance (or the lowest Winkler scores) vary over the eight source markets.

For each source market, the percentages of combined forecasts which outperform the corresponding individual component models are further calculated. The results can be found in Tables 4-5. It is noted that all the combined interval forecasts either outperform the best individual models or better than the worst single models across the eight source markets in most cases, indicating that generally interval combination helps to avoid the worst forecasts. This technique is especially useful when one has no knowledge about the performance of individual forecasting models. To further enhance this conclusion, we examine and compare the worst forecasts (i.e., the highest Winkler scores) for single and combination models.

The part of "highest Winkler scores of single and combination models" in Tables 4-5 reports the highest Winkler scores for single models and seven combination models. For example in Table 4, the figure in the second row of this part and the second column is 0.392, which means that among the 28 cases producing all of the possible combinations based on two models, the highest Winkler score is 0.392. According to Tables 4-5, majority of the combination forecasts (from two-model combinations to eight-model combinations for all source markets and at both confidence levels and both forecasting horizons) outperform the single models when the highest Winkler scores (the worst performance) are considered, with only a few exceptions noted with asterisks in the tables. In practice, it is important to avoid the risk of severe forecast failures. This finding shows that combination forecasting is an effective method for avoiding such failures.

After comparing the highest Winkler scores between the single-model and combination forecasts, a follow-up question arises: does the forecast performance improve on average? In other words, if there is a pool of models available, will the performance expectation be the same between choosing a single model and combining models? We calculate and compare the performance expectation measured using the average Winkler scores for both single and combination models. The results are shown in the bottom of Tables 4 and 5. As expected, the average Winkler scores for majority of the combination cases (from the two- to eight-model combinations) are lower than the corresponding average of the single models' Winkler scores when both confidence levels and both forecasting horizons are concerned. We further obtain consistent findings when the medians of Winkler scores are used. They are not reported here due to space constraints. The above findings confirm that the combination technique improves interval forecast performance.

Conclusion and implications

Conclusion and contribution

The key objectives of this study are to investigate how to combine interval forecasts and whether combination forecasting can improve forecasting accuracy when interval forecasts are requested. To answer these questions, this study forecasts the Hong Kong tourism demand of its eight key source markets: mainland China, Taiwan, Japan, Korea, Macao, Singapore, the Philippines, and the US. Eight single models are used to produce initial interval forecasts: the naïve, ES, ARIMA, STS, ADL, VAR, EC and TVP models. Because directly combining two intervals causes problems due to improper confidence levels for the combined interval, the combination of interval forecasts at various forecasting horizons and at a given confidence level is achieved by combining the density functions of the forecasts. Apart from the coverage rate and the width of the interval, the comprehensive measure Winkler score is used to evaluate the forecasting accuracy of tourism demand. This study examines all of the possible combination options from two to eight models.

This study discloses that combination is an effective way to produce accurate interval forecasts. Although combined intervals cannot always outperform the best individual intervals, they generally outperform not only the worst single intervals but also the average forecasting accuracy of the individual intervals involved, which suggests that the combination of interval forecasts can reduce the risk of forecast failures and improve forecasting performance. Tourism practitioners would like to see the emergence of new methodology in tourism demand forecasting, but ideally they are keener on applying robust methods to generate reliable forecasts and avoiding wrong decisions which are based on failed forecasts. The preference to a robust method will be stronger when practitioners aim to predict the trend of a new market with little prior information available. This study provides empirical support for the advantage of applying interval forecast combination in tourism demand forecasting practice. As no single model can outperform the others in all situations, interval combination is a superior alternative because it can produce more accurate interval forecasts.

There are two methodological contributions in this study. First, differing from previous tourism forecasting studies that focus on point forecast combination, this study is the first to examine how to generate more accurate interval forecasts of tourism demand using the combination technique. The findings provide guidelines for tourism forecasting practices in real life. Second, this study introduces a comprehensive measurement, the Winkler score, to assess forecasting intervals. In the tourism literature, only the coverage rate and interval width are used to measure interval forecasting performance. The Winkler score is superior because it jointly considers these two measures by preferring narrow width and imposing a penalty when the real value is not covered by the interval forecasts.

Implications and future research directions

The findings of this study provide useful practical implications. This study highlights that using only one single model to generate

an interval forecast for an uncertain future is risky and that producing interval forecasts from a few single models and combining these intervals are likely to lead to more accurate forecasts. Illustrations are provided for practical real-life applications.

More accurate interval forecasts of tourism demand provide supplemental information beyond point forecasts to support more comprehensive evidence-based decision-making. For instance, accurate interval forecasts help governments with planning and investment on new tourism development projects or infrastructure. Accurate interval forecasts of tourism demand also benefit industry practitioners, including airlines, hotels, restaurants, transportation and tour operators, for their strategy formulation, including pricing, investment, marketing and revenue management. Knowing different intervals of the tourism demand forecasts with different probabilities would help the above decision makers set up alternative scenarios and develop corresponding plans to deal with different levels of uncertainties and risks.

The findings of this study suggest that further research is necessary. First, this study puts equal weights on single models to combine intervals. Future research should explore different weighting methods and their effects on interval forecasting performance. Second, more individual models, such as artificial intelligence-based models, should be included to produce interval forecasts, and further combination, different origin-destination cases and different data frequencies should be examined and compared.

Acknowledgements

The authors would like to acknowledge the financial support of the National Natural Science Foundation of China (Grant No. 71573289).

References

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting, 27*(3), 822–844.
Athanasopoulos, G., Song, H., & Sun, J. A. (2018). Bagging in tourism demand modeling and forecasting. *Journal of Travel Research, 57*(1), 52–68.
Billio, M., Casarin, R., Ravazzolo, F., & van Dijk, H. K. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics, 177*, 213–232.
Chen, J. L., Li, G., Wu, D. C., & Shen, S. (2019). Forecasting seasonal tourism demand using a multiseries structural time series method. *Journal of Travel Research, 58*(1), 92–103.
Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*, 559–583.
Cortés-Jiménez, I., & Blake, A. (2011). Tourism demand modeling by purpose of visit and nationality. *Journal of Travel Research, 50*(4), 408–416.
Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica, 55*(2), 251–276.
Garratt, A., Lee, K., Pesaran, M. H., & Shin, Y. (2003). Forecast uncertainties in macroeconometric modeling: An application to the UK economy. *Journal of the American Statistical Association, 98*, 829–838.
Garratt, A., Mitchell, J., Vahey, S. P., & Wakerly, E. C. (2011). Real-time inflation forecast densities from ensemble Phillips curves. *The North American Journal of Economics and Finance, 22*, 77–87.
Granger, C. W. J., & Jeon, Y. (2004). Thick modelling. *Economic Modelling, 21*(2), 323–343.
Hall, S. G., & Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting, 23*(1), 1–13.
Harvey, A. C. (1989). *Forecasting, structural time series models and Kalman filter*. Cambridge: Cambridge University Press.
Harvey, A. C., & Jaeger, A. (1993). Detrending, stylized facts and the business-cycle. *Journal of Applied Econometrics, 8*(3), 231–247.
Hendry, D. F. (1986). Empirical modeling in dynamic econometrics. *Applied Mathematics and Computation, 20*, 201–236.
Hendry, D. F., & Clements, M. P. (2004). Pooling of forecasts. *The Econometrics Journal, 7*, 1–31.
Hong, T., & Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting, 32*, 914–938.
Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting, 18*(3), 439–454.
Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, 82*(1), 35–45.
Kascha, C., & Ravazzolo, F. (2012). Combining inflation density forecasts. *Journal of Forecasting, 29*, 231–250.
Kim, J. H., Song, H., & Wong, K. K. F. (2010). Bias-corrected bootstrap prediction intervals for autoregressive model: New alternatives with applications to tourism forecasting. *Journal of Forecasting, 29*(7), 655–672.
Kim, J. H., Wong, K., Athanasopoulos, G., & Liu, S. (2011). Beyond point forecasting: Evaluation of alternative prediction intervals for tourist arrivals. *International Journal of Forecasting, 27*(3), 887–901.
Li, G., Song, H., & Witt, S. F. (2005). Recent developments in econometric modeling and forecasting. *Journal of Travel Research, 44*, 82–99.
Li, G., & Wu, D. C. (2019). Introduction to the special focus: Tourism forecasting – New trends and issues. *Tourism Economics*. https://doi.org/10.1177/1354816618816809.
Liu, B., Nowotarski, J., Hong, T., & Weron, R. (2017). Probabilistic load forecasting via quantile regression averaging on sister forecasts. *IEEE Transactions on Smart Grid, 8*(2), 730–737.
Nowotarski, J., & Weron, R. (2015). Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics, 30*, 791–803.
Shen, S., Li, G., & Song, H. (2008). An assessment of combining tourism demand forecasts over different time horizons. *Journal of Travel Research, 47*(2), 197–207.
Shen, S. J., Li, G., & Song, H. (2011). Combination forecasts of international tourism demand. *Annals of Tourism Research, 38*(1), 72–89.
Song, H., & Li, G. (2008). Tourism demand modelling and forecasting – A review of recent research. *Tourism Management, 29*(2), 203–220.
Song, H., Witt, S. F., Wong, K. F., & Wu, D. C. (2009). An empirical study of forecast combination in tourism. *Journal of Hotel & Tourism Research, 33*(1), 3–29.
Song, H., Wong, K. F., & Chon, K. S. (2003). Modeling and forecasting the demand for Hong Kong tourism. *International Journal of Hospitality Management, 22*, 435–451.
Tay, A. S., & Wallis, K. F. (2000). Density forecasting: A survey. *Journal of Forecasting, 19*, 235–254.
Timmermann, A. G. (2006). Forecast combinations. *Handbook of economic forecasting. 1. Handbook of economic forecasting* (pp. 135–196).
Vu, J. C., & Turner, L. W. (2006). Regional data forecasting accuracy: The case of Thailand. *Journal of Travel Research, 45*(2), 186–193.
Wallis, K. F. (2005). Combining density and interval forecasts: A modest proposal. *Oxford Bulletin of Economics and Statistics, 67*, 983–994.
Wan, S. K., Song, H., & Ko, D. (2016). Density forecasting for tourism demand. *Annual of Tourism Research, 60*, 27–30.
Winkler, R. L. (1972). A decision-theoretic approach to interval estimation. *Journal of American Statistical Association, 67*(337), 187–191.
Wong, K. K. F., Song, H., Witt, S. F., & Wu, D. C. (2007). Tourism forecasting: To combine or not to combine. *Tourism Management, 28*(4), 1068–1078.
Wu, D. C., Song, H., & Shen, S. (2017). New developments in tourism and hotel demand modeling and forecasting. *International Journal of Contemporary Hospitality Management, 29*(1), 507–529.

Gang Li is professor of tourism economics in the School of Hospitality and Tourism Management at the University of Surrey in the UK. His research focuses on

economic analysis of tourism demand.

Doris Chenguang Wu is an associate professor in the Business School at the Sun Yat-sen University, China. Her research interests include tourism demand forecasting and tourist behaviour.

Menglin Zhou is a postgraduate student in the Department of Statistics at the University of British Columbia in Canada. Her research interests include high dimensional data analysis and tourism demand forecasting.

Anyu Liu is a lecturer in hospitality in the School of Hospitality and Tourism Management at the University of Surrey in the UK. His research focuses on applied economics in tourism and hospitality, tourism and hotel demand modelling and forecasting and big data analysis in tourism and hospitality.