Annals of Tourism Research 75 (2019) 379–392

Contents lists available at ScienceDirect

Annals of Tourism Research

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

Invited Article

Density tourism demand forecasting revisited

Haiyan Song[a,*], Long Wen[b,*], Chang Liu[b]

[a] School of Hotel and Tourism Management, The Hong Kong Polytechnic University, Hun Hom, Hong Kong
[b] School of Economics, University of Nottingham Ningbo China, Ningbo, PR China

ARTICLE INFO

Associate editor: Gang Li

Keywords:
Tourism demand
Density forecasts
Scoring rules
Bootstrap

ABSTRACT

This study used scoring rules to evaluate density forecasts generated by different time-series models. Based on quarterly tourist arrivals to Hong Kong from ten source markets, the empirical results suggest that density forecasts perform better than point forecasts. The seasonal autoregressive integrated moving average (SARIMA) model was found to perform best among the competing models. The innovation state space models for exponential smoothing and the structural time-series models were significantly outperformed by the SARIMA model. Bootstrapping improved the density forecasts, but only over short time horizons.

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

Introduction

Tourism forecasting is a well-established research area that attracts a great deal of attention from both academics and practitioners. Private-sector actors need accurate tourism forecasts to make managerial decisions on matters such as pricing and operation strategies. Tourism forecasts also help destination governments to formulate policies and allocate resources more efficiently. A wide range of methods have been applied in tourism forecasting, with the most commonly adopted methods involving time-series models, econometric approaches, and AI-based approaches (Wu, Song, & Shen, 2017). However, forecasting results are often dependent on data characteristics and study features, such as demand measures, data frequency, origin/destination pairs, study sample size, or forecast horizons (Peng, Song, & Crouch, 2014). Song and Li (2008) concluded that no single model could outperform other models consistently in all situations.

It has been widely observed that most previous studies in this field have focused on point forecasting. A point forecast is a number that predicts an unknown future value. One of the problems with point forecasts is that they provide no information concerning the degree of uncertainty involved (Kim, Wong, Athanasopoulos, & Liu, 2011). A comparison of alternative point forecasts may raise the question of reliability unless the levels of uncertainty can be assessed. Wan, Song, and Ko (2016) gave examples of the deficiency of point forecasts of tourism demand when there are varying levels of uncertainty. Therefore, measures of forecast uncertainty can provide important information to decision makers. Interval forecasting is able to convey forecast uncertainty by producing a range of possible future outcomes, given a prescribed level of confidence. To date, however, very few studies have evaluated the performance of interval forecasts in the context of tourism demand. Kim et al. (2011) compared the performances of various interval forecasts of tourist arrivals to Hong Kong and Australia. Mean coverage rates and widths were evaluated for several time-series models. All of the models produced satisfactory results, but models based on a bias-corrected bootstrap performed the best in general. Athanasopoulos, Hyndman, Song, and Wu (2011) evaluated the forecast coverage probabilities of several time-series models used to generate forecast intervals of monthly, quarterly and yearly tourism demand data. They found that the autoregressive integrated moving average (ARIMA) model tended to overestimate the coverage probabilities of the

[*] Corresponding authors.
E-mail addresses: haiyan.song@polyu.edu.hk (H. Song), long.wen@nottingham.edu.cn (L. Wen), chang.liu@nottingham.edu.cn (C. Liu).

https://doi.org/10.1016/j.annals.2018.12.019
Received 12 September 2018; Received in revised form 16 December 2018; Accepted 19 December 2018
0160-7383/ © 2018 Elsevier Ltd. All rights reserved.

forecast intervals, but Forecast Pro and exponential smoothing with trend and seasonality (Hyndman, Koehler, Ord, & Snyder, 2008) produced coverage probabilities that were very close to the nominal rates for monthly and quarterly data. They also found that there was an increased tendency to overestimate coverage probabilities with lower-frequency data.

As interval forecasts can explicitly take uncertainties into account, they can be more informative than point forecasts. However, interval forecasts provide no information about what the tails look like, or about the probabilities of given events. A density forecast for a random variable at some future time is an estimate of the probability distribution for the possible future values of that variable. Such forecasts provide a complete probabilistic description of the possible future realisations of a variable. Both point and interval forecasts can be derived directly from density forecasts.

Density forecasts have the advantage of allowing users to derive probability forecasts for any event they are interested in. This capability is useful for those who need tourism demand forecasts and often base their decisions on the probability of a specific event occurring. For example, a government official may be interested in the probability of tourist arrivals exceeding a certain threshold at some point in the future, at which point more investment will be needed for tourism infrastructure. Hotel managers need forecasts of when guest arrivals are likely to exceed certain thresholds, so that extra staff can be hired in advance. Providing density forecasts can enable both destination governments and practitioners in the tourism industry to allocate limited resources more effectively and to efficiently meet the needs of probable situations. However, only one study, a limited analysis by Wan et al. (2016), has investigated the accuracy of density forecasts for tourism demand. That study highlighted the importance of density forecasts and used the probability integral transform (PIT) to evaluate distributional assumptions in tourism demand on the basis of an autoregressive distributed lag model. The PIT is an absolute evaluation method that assesses the statistical compatibility between density forecasts and actual observations. However, data-generating processes are difficult to identify, and an absolute test for density forecasts is of limited use if it rejects or accepts all models in a comparison. Hence, comparative evaluation methods are needed to determine which model results in better density forecasts. Proper scoring rules can provide comparative assessments of density forecasts, and such rules can be used for model selection. Some scoring rules also enable the direct comparison of point and density forecasts. This study revisits the work of Wan et al. (2016) and is the first attempt to comprehensively evaluate and compare the point and density forecasts of a number of alternative models.

We focus on the inbound tourism demand in Hong Kong from ten major source markets: China, Taiwan, Korea, the United States (USA), Japan, Macau, the Philippines, Singapore, Thailand and Australia. We apply a quarterly time series of tourist arrivals to Hong Kong from these inbound markets.

Six time-series models are used to generate the density forecasts. These models are (1) an autoregressive (AR) model, (2) a bias-corrected bootstrap version of the AR model (Kim, 2004), (3) a seasonal autoregressive integrated moving average (SARIMA) model, (4) an innovation state space model for exponential smoothing with normal errors, (5) an innovation state space model for exponential smoothing with bootstrapped errors (Hyndman et al., 2008) and (6) a structural time-series model (Harvey, 1989). These time-series models were also used by Kim et al. (2011) to forecast tourist arrival intervals. Moreover, all of these models are capable of not only generating interval forecasts, but also of making density forecasts. We adopt a fully automated procedure to select and estimate these models.

The remainder of the study is organized as follows. The next section gives a description of the density forecasts and how they are evaluated. Section “Data and methodology” presents the data and methodology. The results are provided in Section “Results”, and the final section offers the study’s conclusions.

Overview of density forecasting

Density forecasts (probabilistic forecasts) take the form of predictive probability distributions of future quantities or events. Such forecasts offer a complete description of future uncertainty, and thus are more useful to decision makers than point forecasts.

The following example illustrates the differences between point, interval and density forecasts. Assuming a density forecast of the growth rate of tourism demand has a normal distribution with mean 0.1 and standard deviation 0.1, the point, 90% interval and density forecasts are as shown in Fig. 1. A point forecast is often calculated as the mean of the distribution, which is equal to 0.1 in this case. A 90% interval forecast can be constructed by using the 90% confidence interval of the density forecast; if the interval

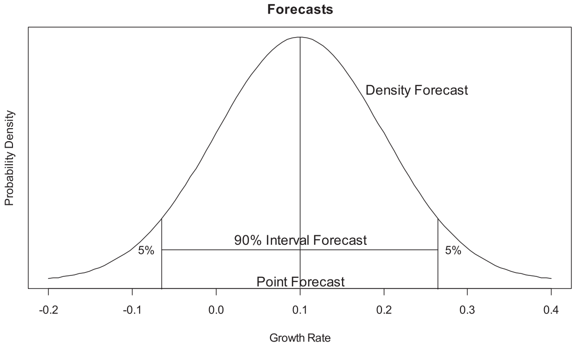

Fig. 1. Point, interval and density forecasts.

forecast is correctly specified, it is expected to cover the future realisations with a probability of 0.9. Finally, the normal distribution itself is the density forecast. It is thus clear that all of the information provided by point and interval forecasts is contained in the density forecast.

Density forecasts have been used for a growing range of purposes, such as making climate and weather forecasts (Leutbecher & Palmer, 2008; Palmer, 2012), population projections (Raftery, Alkema, & Gerland, 2014), epidemiological projections (Alkema, Raftery, & Clark, 2007), presidential election predictions (Montgomery, Hollenbach, & Ward, 2012) and healthcare performance predictions (Jones & Spiegelhalter, 2012).

Density forecasting is also attracting increasing attention in the fields of finance and macroeconomics (Tay & Wallis, 2000). In finance, the most salient example of using density forecasts is in the area of risk management. Density forecasts of changes in portfolio values are needed to generate risk measures such as value-at-risk (VaR), which involves the nth percentile of the distribution. In VaR predictions, the values of customised portfolios are forecast to lose amounts equal to or more than their VaR over specified holding periods, with a probability of $n/100$. The earliest application of density forecasts in macroeconomics was in 1968, when a quarterly survey of macroeconomic forecasters, now known as the Survey of Professional Forecasters, was initiated in the United States (Croushore, 1993). The Bank of England started issuing density forecasts for inflation rates in February 1996, and for gross domestic product in November 1997. These predictions were presented using fan charts. Since then, many countries have started producing density forecasts, including Australia, Brazil, Canada, Norway, the Philippines, South Africa, Thailand and Turkey (Hammond, 2012).

As density forecasts proliferate, the need to evaluate the forecasts is also growing rapidly. The goal of density forecasting is to maximise the sharpness of the predictive distributions, subject to calibration (Gneiting, Balabdaoui, & Raftery, 2007). The notion of calibration refers to the statistical compatibility of density forecasts and their realisations. To demonstrate statistical compatibility, the realisations should resemble random draws from the density forecasts. Sharpness involves the concentration of the density forecasts, and this feature can be assessed by the widths of the associated prediction intervals. Shorter widths indicate better density forecasts, given that the empirical coverage rate is statistically no different from the nominal coverage rate.

Diebold, Gunther, and Tay (1998) first suggested using a histogram-based evaluation technique to check the uniformity of a PIT. The PIT can be calculated by $F(y)$, where F is the cumulative distribution function (CDF) of the density forecast and y is the realised value. For a correctly specified model, the PIT is *i.i.d.* and uniformly distributed on (0,1). With n_b equally sized non-overlapping bins, the model is correctly specified if the histograms are close to flat and there is roughly $h = 1/n_b$ of PIT in each bin with estimated variance $s_{pit}^2 = h(1-h)/n_f$, where n_f is the number of forecasts. The 99% confidence interval for the relative frequencies $\hat{h}$ is therefore $h \pm 2.58 s_{pit}$.

It has been suggested that more formal tests of the uniformity hypothesis should be conducted (Berkowitz, 2001; Hong & White, 2005). However, absolute tests cannot be used for model comparisons if all models produce density forecasts that are calibrated correctly. As an example, histograms of the PIT with a forecast horizon of 1 for the source market China constructed using the six models compared in this study are shown in Fig. 2.

AR, BOOT-AR, SARIMA, ISS1, ISS2 and BSM denote models (1)–(6), respectively (see Section “Models” for details). With $n_b = 5$, the solid line and the two dashed lines represent the mean ($h = 0.2$) and the 99% confidence interval, respectively. It can be seen that all bars of most of the models (AR model has some exceptions) lie inside the boundaries, which suggests that the density forecasts generated by most models are appropriate based on the absolute test using PIT (the *i.i.d.* assumption has also been tested to be valid using the procedure in Diebold et al. (1998)). In cases like this, the absolute test cannot be used for model comparisons and comparative evaluations are needed.

For comparative evaluations of density forecasts, scoring rules can be used instead of the absolute test to help determine the best model. Proper scoring rules provide summary measures of density forecasts and allow for joint assessments of calibration and sharpness (Gneiting & Katzfuss, 2014). Therefore, it is important that the scoring rule is appropriate for comparative evaluations and ranking of the density forecasts of interest. For a detailed review of the mathematical properties of proper scoring rules, see Gneiting and Raftery (2007).

A scoring rule assigns a numerical value $S(F, y)$ to each paired set of a forecast F and a realised value y. Generally, scores are taken to be negatively oriented penalties that the forecasters hope to minimise. Therefore, a lower score indicates a better forecast. In practice, models are ranked by mean scores over the test dataset, and these scores are then used for forecasting comparisons.

The logarithmic score (LogS) is one of the most popular proper scoring rules (Good, 1952). This rule is defined as

$$\mathrm{LogS}(F, y) = -\log f(y), \tag{1}$$

where f is the probability density function (PDF) of F. The LogS identifies the model that, on average, has the higher probability of predicting the realised values.

Another popular proper scoring rule is the continuous rank probability score (CRPS) (Matheson & Winkler, 1976). This rule is defined as

$$\mathrm{CRPS}(F, y) = \int (F(z) - 1\{y \le z\})^2 \mathrm{d}z, \tag{2}$$

where F is the CDF of the density forecast, and $1\{y \le z\})$ is an indicator function, which is 1 if $y \le z$, and 0 otherwise. The CRPS measures the average absolute distance between F and the empirical CDF of y, which is just a step function at y.

For some density forecasts, the predictive distributions are in the form of simulated samples rather than analytical formats. For example, in weather forecasting applications, the density forecasts consist of simulated sample values that are generated by numerical weather prediction models with differing model physics and initial conditions. The simulated samples must be converted into

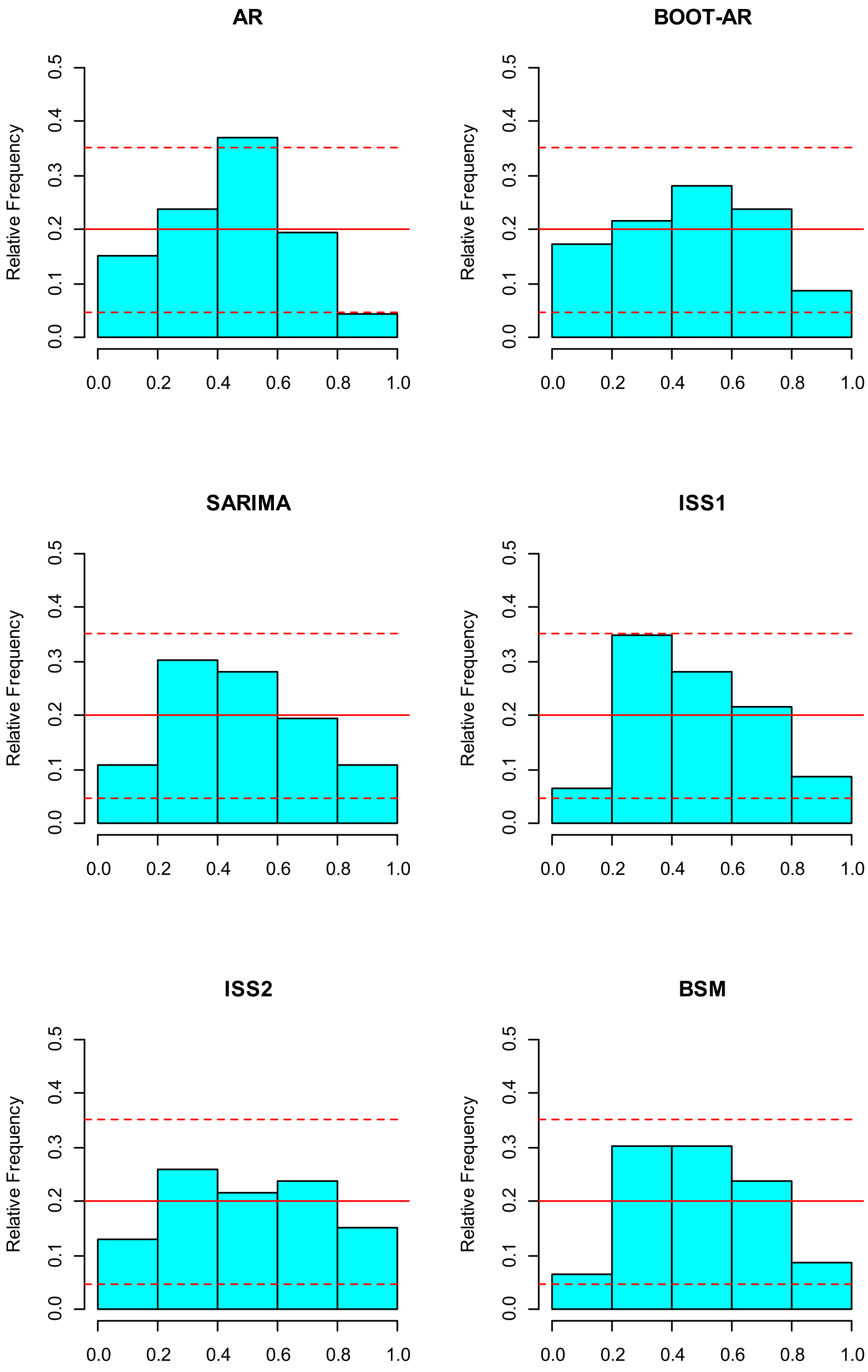

Fig. 2. Histograms of the PIT of the six models used in this study.

distributions with closed-form expressions before the scores are calculated. To obtain the CRPS, an empirical CDF from a simulated sample can be calculated and used as the predictive distribution. If F is a point forecast, then the empirical CDF becomes a step function at the value of the point forecast, and the CRPS is reduced to the absolute error. Therefore, the CRPS offers a direct way of comparing point and density forecasts. Unlike the CRPS, the calculation of the LogS requires a PDF. Kernel density estimation can be used to estimate the PDF from a simulated sample. However, it is only valid under stringent theoretical assumptions. Using the LogS with a kernel density estimation tends to produce fragile results (Krüger, Lerch, Thorarinsdottir, & Gneiting, 2016). Therefore, in this study, the LogS is calculated only for density forecasts with analytical forms.

The absolute tests for density forecasts are unable to differentiate the qualities of the forecasts generated by different models when the PITs all lie within the confidence interval of the uniform distribution. This study first introduces proper scoring rules (LogS and CRPS) for the comparative evaluations of density forecasts in the context of tourism demand.

Data and methodology

Data

In this study, quarterly tourist arrivals to Hong Kong from ten major origin countries in the period Q1 1995 to Q2 2017 were collected from PartnerNet (https://partnernet.hktb.com), the Hong Kong Tourism Board's B2B website for sharing tourism news. These data indicate that there was a large increase in the number of visitors from Macau in Q4 1995, because Macanese visitors arriving via the Macau Ferry Terminal and the China-HK Ferry Terminal were included from 9 November 1995. Therefore, only data from Q1 1996 onwards are used in the Macau models. Hong Kong is one of the most popular tourism destinations in Asia. The total arrivals in Hong Kong increased from 10.2 million in 1995 to 56.7 million in 2016. Despite the rapid growth during this period, the SARS outbreak in 2003 had a major effect on Hong Kong tourism, causing a 90% decline in total arrivals during the second quarter of that year. Various other tourism-disrupting events also occurred during the study period, such as the Asian Financial Crisis and the handover to the People's Republic of China in 1997. However, these events had relatively little impact compared to the SARS outbreak. The log-transformed arrival series from the ten inbound markets is plotted in Fig. 3. The time series is log-transformed following traditional tourism demand modelling and forecasting studies, such as Song and Witt (2000) and Wan et al. (2016).

A large drop in Q2 2003 can be observed in all of the series plots. Some series also show declines in 1997, but they are relatively smaller. Therefore, the SARS outbreak is smoothed out in this study following Kim et al. (2011), who used a similar dataset and smoothed out this unexpected event by using dummy variables.

For ease in interpreting the density forecasts, the year-on-year growth rates of tourist arrivals are calculated and modelled as in Wan et al. (2016). The outlier caused by the SARS outbreak in Q2 2003 is smoothed out before calculating the year-on-year growth rates, as this event distorts the growth rates in both Q2 2003 and Q2 2004. The *tsoutliers* function in the *R* package *forecast* is used to calculate the replacement of the outlier in Q2 2003 for each of log-transformed arrivals data. This function decomposes the series into seasonal and trend components. Then, the outlier replacements are estimated, using the linear interpolation on the seasonally adjusted series. After this adjustment, year-on-year growth rates (y_t) can be calculated.

The adjusted year-on-year growth rates of tourist arrivals are calculated over the period of Q1 1996 to Q2 2017 (except for Macau, which is calculated from Q1 1997). The data up to Q4 2005 are used for model estimation, and the data for later periods are used as a test dataset for the forecasting comparison.

Models

The specifications of the six models that are used to generate density forecasts are presented in the following subsections.

AR model with bias-corrected bootstrap

An AR(p) model can be expressed as follows:

$$y_t = \delta + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + u_t, \tag{3}$$

where $\{y_t\}_{t=1}^{n}$ is the observed time series, p is the order of autoregression, and u_t is $i.\ i.\ d.$ with mean zero and fixed variance. $(\delta, \phi_1, \cdots, \phi_p)$ are the parameters that need to be estimated. A fully automated procedure is adopted to select the lag order, and to decide whether non-zero mean and trend terms should be included, based on the corrected Akaike Information Criterion (AIC) (Hyndman & Khandakar, 2008). The parameters are estimated using maximum likelihood estimation (MLE). After estimation, the point forecasts for y_{n+1} at time n can be generated using past p observations, and forecasts for longer horizons can be generated recursively. Finally, density forecasts of the AR model can be constructed based on normal approximations of the error terms.

The bootstrap method is commonly found to generate better prediction intervals because it is robust to non-normality and it accounts for uncertainty in the parameter estimates. Bias-corrected bootstrap methods have the advantage of overcoming the biases of parameter estimators in small samples, and these methods are found to produce better interval forecasts (Kim, 2001, 2004). The bias-corrected bootstrap method developed by Kim (2004) is used in this study to investigate whether a bootstrap can produce better density forecasts for the AR model. A simple description of this method is given here. After selecting the lag order using the same procedure as is used for a normal AR model, the parameters are estimated using the least squares method. The bias of the estimated parameters is assessed using the analytical formula proposed by Shaman and Stine (1988). Bias-corrected estimators can be derived as $\left(\hat{\delta}^c, \hat{\phi}_1^c, \cdots, \hat{\phi}_p^c\right)$. After bias-correction, the residuals can be calculated as $\{e_t^c\}_{t=p+1}^{n}$. A backward AR form can then be used to generate an artificial dataset $\{y_t^*\}_{t=1}^{n-p}$ recursively, as follows:

$$y_t^* = \hat{\delta}^c + \hat{\phi}_1^c y_{t+1}^* + \cdots + \hat{\phi}_p^c y_{t+p}^* + u_t^*, \tag{4}$$

where the last p values of $\{y_t\}_{t=1}^{n}$ are set as the starting values, and u_t^* is a random draw from $\{e_t^c\}_{t=p+1}^{n}$ with replacement. Model (3) is

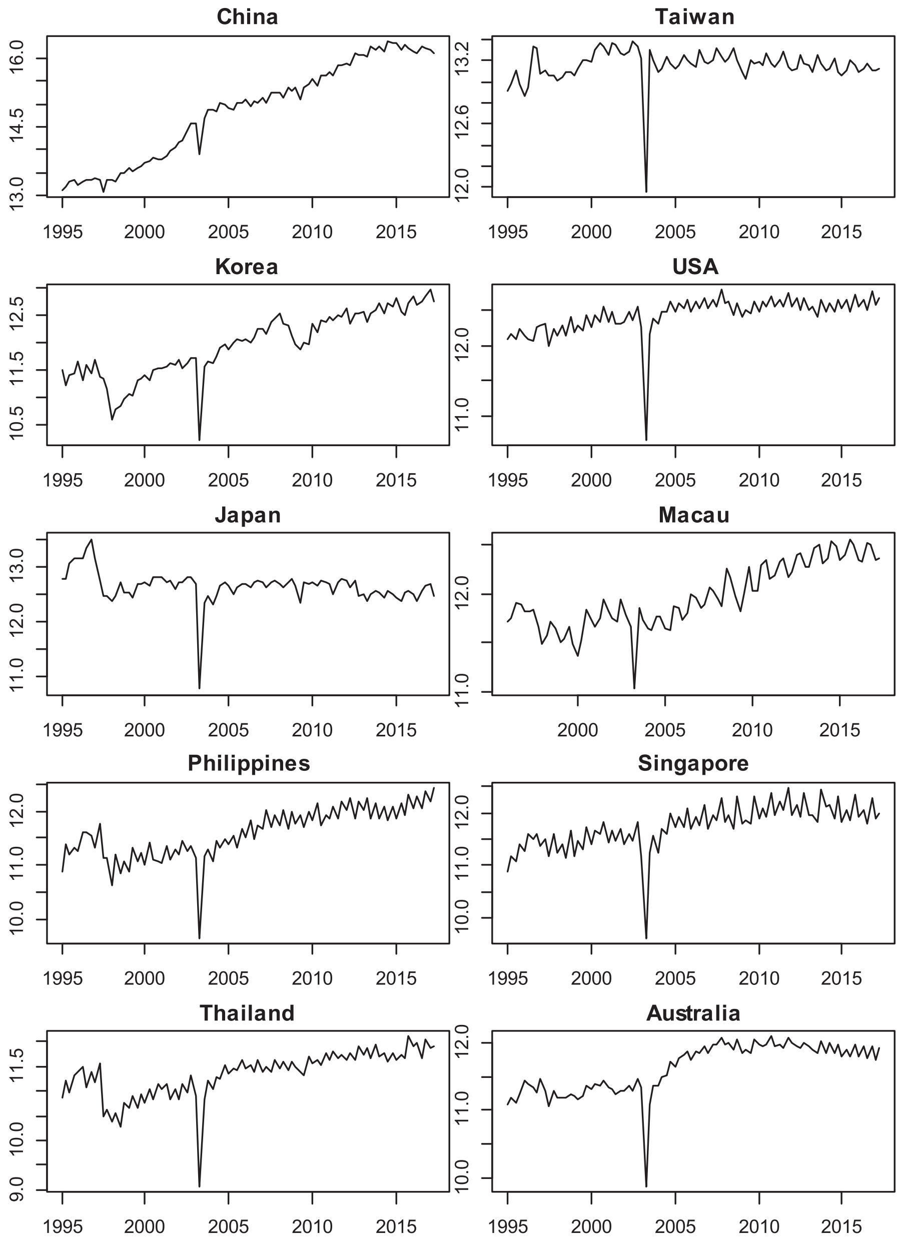

Fig. 3. Time plots of arrival series (natural logarithms).

estimated again, using the generated artificial dataset $\{y_t^*\}_{t=1}^n$. The same bias-correction procedure is used, and the new bias-corrected estimators can be calculated as $\left(\hat{\delta}^{*c}, \hat{\phi}_1^{*c}, \cdots, \hat{\phi}_p^{*c}\right)$. A point forecast for y_{n+h} can be generated recursively using this bootstrap replicate, as is done for a normal AR model. The generation of artificial datasets can be repeated many times and the bootstrap distribution for AR forecasts can be produced. This distribution is the density forecast for the bootstrapped AR model.

SARIMA model

The SARIMA model was first proposed by Box and Jenkins (1970). In addition to incorporating autoregressive terms, this model can also incorporate differencing and moving average components. The general form of SARIMA(p,d,q)(P,D,Q)$_m$ can be expressed as

$$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d y_t = \delta + \Theta(B^m)\theta(B)u_t, \tag{5}$$

where u_t is a white noise process with mean zero and fixed variance, m is the order of periodicity, B is the backshift operator, and$(1-B^m)^D$ and $(1-B)^d$ are the seasonal and annual differencing order terms. $\Phi(x)$ and $\Theta(x)$ are polynomials of orders P and Q, respectively. $\phi(x)$ and $\theta(x)$ are polynomials of orders p and q, respectively. An automatic procedure is used to select the orders (Hyndman & Khandakar, 2008). The seasonal differencing order D is chosen first, using the OCSB test (Osborn, Chui, Smith, & Birchenhall, 1988). Then, the order of non-seasonal differencing d is chosen, based on the KPSS unit-root test (Kwiatkowski, Phillips, Schmidt, & Shin, 1992). Finally, the order of the autoregressive and moving average terms p, q, P, and Q is determined, based on a corrected AIC, using a step-wise procedure to traverse the model space. The parameters are then estimated using MLE. As with the AR model, the point forecasts can be first generated recursively, and then the density forecasts can be generated using a normal approximation of the error terms for the SARIMA.

Innovation state space models for exponential smoothing

Traditional exponential smoothing methods, such as simple smoothing, or the Holt and Holt–Winters seasonal exponential smoothing methods, can only produce point forecasts. Innovation state space (ISS) models for exponential smoothing provide a framework for computing density forecasts (Hyndman et al., 2008). Unlike the AR and SARIMA models, which are widely used in tourism demand forecasting, ISS models are little used. However, ISS models have been found to perform well for generating point or interval forecasts in the tourism demand context (Athanasopoulos & Hyndman, 2008; Athanasopoulos et al., 2011; Kim et al., 2011). The general form of an ISS model can be expressed by a measurement equation and a state equation, as follows:

$$y_t = w(\boldsymbol{x}_{t-1}) + r(\boldsymbol{x}_{t-1})\varepsilon_t; \tag{6}$$

$$\boldsymbol{x}_t = f(\boldsymbol{x}_{t-1}) + g(\boldsymbol{x}_{t-1})\varepsilon_t, \tag{7}$$

where $\boldsymbol{x}_t = (l_t, b_t, s_t, s_{t-1}, \cdots, s_{t-m-1})'$ is the state vector; l_t, b_t and s_t denote the level, slope and seasonal components at time t, respectively; m is the order of periodicity; and ε_t is Gaussian white noise with zero mean and fixed variance. The term 'innovation' in the model name reflects the fact that the same random error process ε_t applies to both the measurement and the state equations. The errors involved can be either additive or multiplicative. Models with multiplicative errors are not numerically stable when the data have zero or negative values (Hyndman & Khandakar, 2008).

As this study focuses on predicting the growth rate of tourism demand (which can have a negative value), only models with additive errors are used. An automatic model selection method can be used to choose the forms of the trend and the seasonal components, based on corrected AIC. The parameters are estimated using MLE. Point forecasts can be obtained from ISS models by iterating the equations and by setting $\varepsilon_t = 0$ for $t > n$. Density forecasts can then be generated by using a normal approximation. A bootstrap approach can also be applied by simulating future sample paths with the use of re-sampled errors. Both methods are used to generate density forecasts for the ISS models in this study.

Structural time-series model

A structural time-series model can be written in a state space form, based on a decomposition of the series into a number of unobservable components. The basic structural model (BSM) proposed by Harvey (1989) can be written as

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \tag{8}$$

where μ_t is the trend component, γ_t is the seasonal component and ε_t is Gaussian white noise with zero mean and variance σ_ε^2. The trend and seasonal components can be expressed as

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t, \tag{9}$$

$$\beta_t = \beta_{t-1} + \zeta_t, \tag{10}$$

$$\gamma_t = -\sum_{j=1}^{m-1} \gamma_{t-j} + w_t, \tag{11}$$

where $\eta_t \sim NID(0, \sigma_\eta^2)$, $\zeta_t \sim NID(0, \sigma_\zeta^2)$, and $w_t \sim NID(0, \sigma_w^2)$. When $\sigma_w^2 = 0$, the seasonal pattern becomes deterministic. This pattern is estimated using the *StructTS* function in the *R* package *stats*. The variances and unobservable components can be estimated using MLE. The Kalman filter is then used for the model, which is expressed in state space form to generate point forecasts. Density forecasts of the BSM model can then be computed by assuming the normality of error terms (Harvey, 1989).

Both the ISS and the BSM models are expressed in state space forms, but the error terms are specified differently. The ISS model has the same error term in both the measurement and the state equation, but the BSM model has different error terms in each equation. The BSM model is well-known in the tourism demand forecasting literature, and has been found to produce satisfactory results (Greenidge, 2001; Vu & Turner, 2006).

Results

In this section, we first compare the point and density forecasts of the AR, SARIMA, ISS and BSM models (with analytical forms). Then the density forecasts generated from the alternative time-series models are ranked, including those from the AR model, the bias-corrected bootstrapped AR model (BOOT-AR), the SARIMA model, the ISS model for exponential smoothing with normal

Table 1
Combinations of models, forecast horizons and countries in which the point forecasts outperform the density forecasts.

Horizon	ISS			Horizon	BSM		
	6 quarters	7 quarters	8 quarters		6 quarters	7 quarters	8 quarters
Taiwan		√	√	Taiwan		√	√
Japan	√	√	√	Japan	√	√	√
Macau				Macau		√	√
Thailand				Thailand			√
Australia		√	√	Australia		√	√

Note: √ indicates better point forecasts than density forecasts.

approximated errors (ISS1), the ISS model with bootstrapped errors (ISS2) and the BSM. The model selection and estimations are conducted automatically using the *R* packages *forecast* and *BootPR*.

The performances of the forecasts are evaluated in a purely empirical setting. The data up to Q4 2005 are used for model estimation, and ex-ante 1- to 8-steps-ahead forecasts are generated over the hold-out period of Q1 2006 to Q2 2017. A dynamic forecasting procedure is adopted. The models are first estimated over the period Q1 1996 to Q4 2005 (except for Macau, Q1 1997 to Q4 2005), and then the forecasts are generated for each inbound market and for each forecast horizon of 1–8 steps ahead. Forecasts that have horizons of longer than 1 step are generated recursively. Next, the models are re-estimated using the data up to Q1 2006, Q2 2006, … and Q1 2017, and the forecasts are generated again at each round. In total, 46 forecasts of 1-step-ahead, 45 2-steps-ahead, … and 39 8-steps-ahead projections are obtained for each of the ten inbound markets.

Point forecasts and density forecasts

The CDF for a point forecast is essentially a step function at the value of the forecasted point. Therefore, the CRPS can be used to directly compare the accuracy of point and density forecasts. Point and density forecasts (with normal approximated errors) are generated from the AR, SARIMA, ISS and BSM models. Then comparisons of the models are conducted over all of the forecast horizons and countries. Out of $4 \times 8 \times 10 = 320$ comparisons (4 models, 8 forecast horizons, 10 countries), there are only 17 cases in which the point forecasts perform better than the density forecasts. These cases are summarised in Table 1.

The proportion of pairs with better point forecasts is very small (about 5%). Only the ISS and BSM models have better point forecasts for longer forecast horizons. To see the overall performance of the point and density forecasts, the CRPSs are further averaged over all countries. The results are shown in Table 2.

It is clear that density forecasts outperform point forecasts, and that this outcome is consistent for all of the models and forecast horizons. To see whether this level of outperformance is statistically significant, the Diebold–Mariano (DM) test (proposed by Diebold and Mariano (1995)) is used. A one-tailed test is conducted to determine if the density forecasts significantly outperform the point forecasts.

Table 3 shows the results of the DM test. Almost all of the density forecasts significantly outperform the point forecasts, except for the BSM at the forecast horizon of 8 steps ahead. For the AR and SARIMA models, all of the density forecasts outperform the point forecasts at the 1% significance level. For the ISS and BSM models, this outperformance becomes less obvious for longer forecast horizons.

Rankings of density forecasts

The rankings of the density forecasts are conducted with respect to the CRPS and LogS. The results of the 1-step-ahead density forecasts, as evaluated by the CRPS, are shown in Table 4. The SARIMA performs best for six of the ten origin countries, followed by

Table 2
Comparison of point and density forecasts, as measured by their CRPSs.

Horizon	AR		SARIMA		ISS		BSM	
	Density	Point	Density	Point	Density	Point	Density	Point
1 quarter	**0.05056**	0.06579	**0.04650**	0.06207	**0.05367**	0.06981	**0.05492**	0.07144
2 quarters	**0.06231**	0.08094	**0.06061**	0.08054	**0.07265**	0.09432	**0.07463**	0.09742
3 quarters	**0.06746**	0.08650	**0.06786**	0.09118	**0.08602**	0.10728	**0.08923**	0.11109
4 quarters	**0.07147**	0.09256	**0.07267**	0.09822	**0.09901**	0.12359	**0.10430**	0.12926
5 quarters	**0.07127**	0.09241	**0.07185**	0.09630	**0.10232**	0.12591	**0.10912**	0.13387
6 quarters	**0.07204**	0.09373	**0.07190**	0.09617	**0.10592**	0.12521	**0.11467**	0.13497
7 quarters	**0.07251**	0.09416	**0.07161**	0.09535	**0.10925**	0.12465	**0.12044**	0.13683
8 quarters	**0.07263**	0.09412	**0.07162**	0.09462	**0.11187**	0.12163	**0.12477**	0.13554

Note: Bold indicates better performance.

Table 3
Diebold–Mariano tests for the differences between point and density forecasts.

Horizon	AR	SARIMA	ISS	BSM
1 quarter	-11.56^{***}	-14.00^{***}	-11.63^{***}	-11.52^{***}
2 quarters	-9.22^{***}	-10.84^{***}	-9.68^{***}	-9.89^{***}
3 quarters	-7.65^{***}	-10.01^{***}	-6.76^{***}	-6.28^{***}
4 quarters	-7.07^{***}	-8.60^{***}	-5.90^{***}	-5.43^{***}
5 quarters	-6.40^{***}	-7.84^{***}	-4.71^{***}	-4.37^{***}
6 quarters	-6.37^{***}	-7.56^{***}	-3.22^{***}	-2.87^{***}
7 quarters	-6.11^{***}	-7.03^{***}	-2.35^{***}	-2.00^{**}
8 quarters	-5.95^{***}	-6.53^{***}	-1.43^{*}	-1.26

Note: *, **, and *** denote significance at the 10%, 5% and 1% levels (one-tailed tests).

Table 4
One-quarter ahead forecasts (CRPS).

	AR	BOOT-AR	SARIMA	ISS1	ISS2	BSM
China	0.04497(3)	0.04320(2)	0.04197(1)	0.04612(5)	0.04522(4)	0.04662(6)
Taiwan	0.02877(3)	0.02740(1)	0.02770(2)	0.03244(5)	0.03064(4)	0.03283(6)
Korea	0.07996(2)	0.08050(3)	0.06586(1)	0.08543(5)	0.08509(4)	0.08995(6)
USA	0.03134(1)	0.03217(2)	0.03672(6)	0.03393(4)	0.03336(3)	0.03444(5)
Japan	0.06257(2)	0.06330(3)	0.04862(1)	0.06889(5)	0.06426(4)	0.06974(6)
Macau	0.05040(5)	0.04963(4)	0.04421(1)	0.04926(3)	0.04869(2)	0.05106(6)
Philippines	0.05167(1)	0.05178(2)	0.05433(3)	0.05550(4)	0.05572(5)	0.05747(6)
Singapore	0.05468(2)	0.05743(3)	0.05146(1)	0.05795(4)	0.05921(6)	0.05864(5)
Thailand	0.06315(3)	0.06308(2)	0.05672(1)	0.06797(5)	0.06684(4)	0.06873(6)
Australia	0.03813(3)	0.03664(1)	0.03726(2)	0.03921(5)	0.03907(4)	0.03972(6)
Average	0.05056(3)	0.05051(2)	0.04648(1)	0.05367(5)	0.05281(4)	0.05492(6)

Note: Figures in parentheses denote rankings.

the AR and the BOOT-AR, each of which perform best for two countries. It can be seen from the last row of Table 1 that when the CRPS is averaged over all countries, the SARIMA model ranks first, performing better than the other models overall. The BOOT-AR ranks second and the AR ranks third. The BOOT-AR generates only slightly better results than the AR. The ISS2 ranks fourth and the ISS1 ranks fifth. These results suggest that a bootstrap improves short-term density forecasting. The BSM performs the worst for most countries, ranking last on average.

Table 5 presents the results of 1-step-ahead density forecasts, as measured by the LogS. The findings reported above are largely supported by the LogS results. The rankings stay the same for the four models with predictive distributions that have analytical forms. In this comparison, the SARIMA model still has the best performance.

Table 6 shows the results of the 4-steps-ahead density forecasts, as evaluated in terms of the CRPS. In these forecasts, both the AR and SARIMA models rank first for four countries, but the AR performs slightly better than the SARIMA, ranking first on average. The ISS1 ranks fourth, and the ISS2 ranks fifth. Unlike in the 1-step-ahead forecasts, a bootstrap fails to improve the density forecast for the 4-steps-ahead forecasts. The LogS results of the 4-steps-ahead forecasts are presented in Table 7. They are in line with the findings shown in Table 6.

Tables 8 and 9 present the results of the 8-steps-ahead forecasts. The SARIMA performs best for most countries and is the best model in terms of both the CRPS and LogS. As with the results for the 4-steps-ahead forecasts, a bootstrap does not improve the performance of the 8-steps-ahead forecasts.

Table 5
One-quarter ahead forecasts (LogS).

	AR	SARIMA	ISS1	BSM
China	−1.01207(2)	−1.09834(1)	−0.99171(3)	−0.97783(4)
Taiwan	−1.45014(2)	−1.51583(1)	−1.33139(3)	−1.31741(4)
Korea	−0.48506(2)	−0.66506(1)	−0.41298(3)	−0.37873(4)
USA	−1.36495(1)	−1.28335(2)	−1.21722(3)	−1.20144(4)
Japan	−0.69869(2)	−0.92754(1)	−0.58372(3)	−0.56980(4)
Macau	−0.81634(4)	−0.84160(3)	−0.89408(1)	−0.85386(2)
Philippines	−0.85686(2)	−0.89316(1)	−0.80113(3)	−0.77432(4)
Singapore	−0.85120(2)	−0.94236(1)	−0.80523(3)	−0.79153(4)
Thailand	−0.69413(2)	−0.80415(1)	−0.60470(3)	−0.59122(4)
Australia	−1.12893(2)	−1.20140(1)	−1.09754(3)	−1.08340(4)
Average	−0.93584(2)	−1.01728(1)	−0.87397(3)	−0.85395(4)

Note: The figures in parentheses denote rankings.

Table 6
Four-quarters ahead forecasts (CRPS).

	AR	BOOT-AR	SARIMA	ISS1	ISS2	BSM
China	0.07165(2)	0.07036(1)	0.07628(4)	0.07664(5)	0.07515(3)	0.07789(6)
Taiwan	0.03815(2)	0.03858(3)	0.03750(1)	0.05989(5)	0.05713(4)	0.06019(6)
Korea	0.13051(2)	0.13702(3)	0.10354(1)	0.18404(4)	0.18635(5)	0.21307(6)
USA	0.04906(1)	0.05029(2)	0.05811(3)	0.06991(4)	0.07041(6)	0.07023(5)
Japan	0.07498(2)	0.07904(3)	0.06598(1)	0.12348(5)	0.12297(4)	0.12446(6)
Macau	0.05468(1)	0.05744(2)	0.07293(5)	0.07259(4)	0.07239(3)	0.08558(6)
Philippines	0.07095(1)	0.07267(2)	0.08042(3)	0.09528(4)	0.09636(5)	0.10089(6)
Singapore	0.07924(1)	0.08348(3)	0.08071(2)	0.10236(4)	0.10508(6)	0.10297(5)
Thailand	0.08673(2)	0.09006(3)	0.07995(1)	0.13400(4)	0.13421(5)	0.13509(6)
Australia	0.05880(2)	0.05751(1)	0.06369(3)	0.07195(5)	0.07089(4)	0.07262(6)
Average	0.07147(1)	0.07364(3)	0.07191(2)	0.09901(4)	0.09910(5)	0.10430(6)

Note: Figures in parentheses denote rankings.

Table 7
Four-quarters ahead forecasts (LogS).

	AR	SARIMA	ISS1	BSM
China	−0.58500(1)	−0.51233(2)	−0.48734(3)	−0.47188(4)
Taiwan	−1.13393(2)	−1.17922(1)	−0.67116(3)	−0.66241(4)
Korea	−0.01721(2)	−0.26013(1)	0.33799(3)	0.42915(4)
USA	−0.97944(1)	−0.84002(2)	−0.51816(3)	−0.51474(4)
Japan	−0.43354(2)	−0.61776(1)	0.04276(3)	0.05303(4)
Macau	−0.84639(1)	−0.42535(3)	−0.50923(2)	−0.33595(4)
Philippines	−0.54423(1)	−0.51112(2)	−0.23738(3)	−0.18882(4)
Singapore	−0.46644(2)	−0.51019(1)	−0.17340(3)	−0.16586(4)
Thailand	−0.31385(2)	−0.40708(1)	0.08091(3)	0.09139(4)
Australia	−0.73387(1)	−0.69971(2)	−0.46540(3)	−0.45742(4)
Average	−0.60539(1)	−0.59629(2)	−0.26004(3)	−0.22235(4)

Note: Figures in parentheses denote rankings.

Table 8
Eight-quarters ahead forecasts (CRPS).

	AR	BOOT-AR	SARIMA	ISS1	ISS2	BSM
China	0.07933(1)	0.08298(2)	0.08598(3)	0.09321(4)	0.09370(5)	0.09557(6)
Taiwan	0.03916(1)	0.03922(2)	0.04035(3)	0.07153(5)	0.06713(4)	0.07219(6)
Korea	0.12048(2)	0.12742(3)	0.11242(1)	0.19036(4)	0.19207(5)	0.24968(6)
USA	0.04929(2)	0.04980(3)	0.04661(1)	0.08703(4)	0.08853(6)	0.08796(5)
Japan	0.08173(2)	0.08679(3)	0.07446(1)	0.13425(4)	0.13498(5)	0.13618(6)
Macau	0.05235(1)	0.05517(2)	0.05985(3)	0.07833(4)	0.07881(5)	0.11696(6)
Philippines	0.07067(2)	0.07449(3)	0.06571(1)	0.11734(4)	0.11754(5)	0.13688(6)
Singapore	0.08163(2)	0.08517(3)	0.07916(1)	0.11920(4)	0.12257(6)	0.12055(5)
Thailand	0.08805(2)	0.09081(3)	0.08731(1)	0.14246(5)	0.14241(4)	0.14487(6)
Australia	0.06366(2)	0.05988(1)	0.06429(3)	0.08494(5)	0.08456(4)	0.08688(6)
Average	0.07263(2)	0.07517(3)	0.07161(1)	0.11187(4)	0.11223(5)	0.12477(6)

Note: Figures in parentheses denote rankings.

To assess the forecasting performance over all countries, the CRPS and the average LogS are summarised in Tables 10 and 11, respectively. From Table 10, we can see that the rankings of the different models stay relatively consistent over different forecast horizons. The SARIMA ranks first overall and performs best for all forecast horizons except for the 4-steps-ahead forecast. The AR and the BOOT-AR generate comparable results and are only outperformed by the SARIMA model. The ISS1 and ISS2 produce similar results, performing slightly better than the BSM. The ISS and BSM models are largely outperformed by the SARIMA, AR and BOOT-AR models, especially for longer forecast horizons.

The BOOT-AR performs better than the AR for 1-step-ahead forecasts and the ISS2 outperforms ISS1 for 1- to 3-steps-ahead forecasts. These results indicate that the bootstrap improves density forecasts only over short forecast horizons.

The LogS results presented in Table 11 are in line with the CRPS rankings. For both of these scoring rules, there is a general trend of increasing score as the forecast horizon extends. This set of results suggests a deterioration of density forecasting performance for longer forecast horizons. To further confirm whether the SARIMA significantly outperforms other models, a one-tailed test is conducted. The results are shown in Tables 12 and 13.

As Table 12 shows, the SARIMA model outperforms the ISS1, ISS2 and BSM models at the 1% significance level for all forecast

Table 9
Eight-quarters ahead forecasts (LogS).

	AR	SARIMA	ISS1	BSM
China	−0.49930(1)	−0.41337(2)	−0.25333(3)	−0.23036(4)
Taiwan	−1.11749(2)	−1.12538(1)	−0.39471(3)	−0.38096(4)
Korea	−0.09347(2)	−0.17775(1)	0.40900(3)	0.51468(4)
USA	−0.96415(2)	−1.00082(1)	−0.22308(3)	−0.21614(4)
Japan	−0.35360(2)	−0.48684(1)	0.28014(3)	0.29585(4)
Macau	−0.87967(1)	−0.76564(2)	−0.38081(3)	−0.07841(4)
Philippines	−0.52808(2)	−0.62005(1)	0.03713(3)	0.13173(4)
Singapore	−0.42851(2)	−0.50609(1)	0.07856(3)	0.09040(4)
Thailand	−0.28337(2)	−0.32038(1)	0.28759(3)	0.30422(4)
Australia	−0.66929(2)	−0.69027(1)	−0.18513(3)	−0.17098(4)
Average	−0.58169(2)	−0.61066(1)	−0.03446(3)	0.02600(4)

Note: Figures in parentheses denote rankings.

Table 10
Average CRPS over all countries of origin.

Horizon	AR	BOOT-AR	SARIMA	ISS1	ISS2	BSM
1 quarter	0.05056(3)	0.05051(2)	0.04648(1)	0.05367(5)	0.05281(4)	0.05492(6)
2 quarters	0.06231(2)	0.06323(3)	0.06019(1)	0.07265(5)	0.07203(4)	0.07463(6)
3 quarters	0.06746(2)	0.06947(3)	0.06744(1)	0.08602(5)	0.08512(4)	0.08923(6)
4 quarters	0.07147(1)	0.07364(3)	0.07191(2)	0.09901(4)	0.09910(5)	0.10430(6)
5 quarters	0.07127(2)	0.07402(3)	0.07113(1)	0.10232(4)	0.10263(5)	0.10912(6)
6 quarters	0.07204(2)	0.07571(3)	0.07140(1)	0.10592(4)	0.10609(5)	0.11467(6)
7 quarters	0.07251(2)	0.07507(3)	0.07136(1)	0.10925(4)	0.11012(5)	0.12044(6)
8 quarters	0.07263(2)	0.07517(3)	0.07161(1)	0.11187(4)	0.11223(5)	0.12477(6)

Note: Figures in parentheses denote rankings.

Table 11
Average LogS over all countries of origin.

Horizon	AR	SARIMA	ISS1	BSM
1 quarter	−0.93584(2)	−1.01728(1)	−0.87397(3)	−0.85395(4)
2 quarters	−0.73701(2)	−0.77975(1)	−0.57934(3)	−0.55635(4)
3 quarters	−0.65065(2)	−0.66362(1)	−0.39838(3)	−0.36732(4)
4 quarters	−0.60539(1)	−0.59629(2)	−0.26004(3)	−0.22235(4)
5 quarters	−0.60363(2)	−0.60553(1)	−0.20076(3)	−0.15620(4)
6 quarters	−0.59353(2)	−0.60628(1)	−0.13799(3)	−0.08884(4)
7 quarters	−0.58622(2)	−0.61276(1)	−0.08338(3)	−0.02553(4)
8 quarters	−0.58169(2)	−0.61066(1)	−0.03446(3)	0.02600(4)

Note: Figures in parentheses denote rankings.

Table 12
Diebold–Mariano tests for the differences (in terms of CRPS) between the SARIMA model and the other models.

Horizon	AR	BOOT-AR	ISS1	ISS2	BSM
1 quarter	-3.28^{***}	-3.05^{***}	-4.87^{***}	-4.06^{***}	-5.52^{***}
2 quarters	-1.17	-1.48^{*}	-4.84^{***}	-4.40^{***}	-5.25^{***}
3 quarters	-0.01	-0.83	-4.95^{***}	-4.55^{***}	-4.96^{***}
4 quarters	0.15	-0.55	-5.14^{***}	-5.04^{***}	-4.66^{***}
5 quarters	-0.06	-0.97	-5.57^{***}	-5.46^{***}	-4.84^{***}
6 quarters	-0.27	-1.49^{*}	-5.93^{***}	-5.81^{***}	-4.67^{***}
7 quarters	-0.59	-1.54^{*}	-6.83^{***}	-6.74^{***}	-4.87^{***}
8 quarters	-0.62	-1.65^{**}	-7.62^{***}	-7.50^{***}	-5.04^{***}

Note: *, ** and *** denote significance at the 10%, 5%, and 1% levels (one-tailed tests).

horizons. The SARIMA also outperforms the AR and the BOOT-AR models at the 1% significance level for 1-step-ahead forecasts. Furthermore, the SARIMA model performs better than the BOOT-AR at the 5% or 10% significance levels for longer forecast horizons (6–8 steps ahead). As in Table 12, in which the results are given in terms of the CRPS, Table 13 shows that when the density forecasts are evaluated by the LogS, the SARIMA again outperforms the AR at the 10% significance level for 2- and 8-steps-ahead forecasts.

The only positive statistic appears in the AR model at the forecast horizon of 4 quarters. This irregularity appears because the AR

Table 13
Diebold–Mariano tests for the differences (in terms of LogS) between the SARIMA model and other models.

Horizon	AR	ISS1	BSM
1 quarter	-3.94^{***}	-5.26^{***}	-5.93^{***}
2 quarters	-1.60^{*}	-6.34^{***}	-6.89^{***}
3 quarters	-0.39	-6.53^{***}	-7.16^{***}
4 quarters	0.23	-6.79^{***}	-7.50^{***}
5 quarters	-0.06	-8.55^{***}	-9.83^{***}
6 quarters	-0.48	-10.81^{***}	-12.24^{***}
7 quarters	-1.24	-12.31^{***}	-13.77^{***}
8 quarters	-1.45^{*}	-12.59^{***}	-13.48^{***}

Note: *, ** and *** denote significance at the 10%, 5%, and 1% levels (one-tailed tests).

model performs slightly better than the SARIMA for 4-steps-ahead forecasts. However, the low absolute value of this statistic indicates that this level of outperformance is not significant.

Discussion

Past studies in tourism demand forecasting have mainly focused on point forecasting. No study has yet investigated comparisons between density forecasts and point forecasts. The CRPS offers us a tool for conducting a direct comparison. Using this tool, it is found that density forecasts significantly outperform point forecasts in most cases, especially for the AR and SARIMA models. For the ISS and BSM models, point forecasts perform better than density forecasts for some countries and over longer forecast horizons. However, when the results are aggregated over all countries, density forecasts significantly outperform point forecasts, except in the case of the BSM model at the forecast horizon of 8 steps ahead. In addition to the increased information that density forecasts can provide, their superior accuracy provides further evidence that density forecasts should be used more often.

The ISS and BSM models are found to produce the least accurate density forecasts. This contradicts the findings of Kim et al. (2011), who found that the ISS and BSM models provide reasonably good interval forecasts for tourism demand. Interval forecasts are often measured in terms of coverage and width. The empirical coverage rate can assess how often realisations fall within the prediction intervals, but it does not indicate where a realisation will land in an interval. Scoring rules evaluate the relationship between the predictive distribution and the realised value. As such, the evaluation results of interval and density forecasts are not directly comparable. However, the wide prediction intervals of the ISS and BSM models can to some extent explain their poor performances in density forecasting. Kim et al. (2011) found that although the ISS and BSM models generate accurate empirical coverage rates, they also generate wider prediction intervals than the other models, especially for longer forecast horizons. Wider prediction intervals indicate that the predicted distributions are not tight and thus they reflect lower degrees of sharpness. As proper scoring rules enable the joint assessment of calibration and sharpness, a low degree of sharpness is likely to result in a poor density forecast.

The SARIMA model is found to generate the best density forecasts and to statistically outperform the ISS and BSM models at the 1% significance level over all time horizons. The SARIMA is known to produce decent interval forecasts with tight prediction intervals (Kim et al., 2011). Therefore, it can be expected that the SARIMA model will generate accurate density forecasts.

As with other forms of forecasting, the accuracy of density forecasts deteriorates as the forecast horizon extends. This is particularly the case for bootstrapped models. Bootstrapping is found to improve the density forecasts only over short horizons. This finding is in line with the results of past studies on tourism demand-interval forecasting (Kim, Song, & Wong, 2010; Kim et al., 2011). In general, the performance of bootstrapped models deteriorates as the forecasting horizon increases.

Conclusions

Point forecasts provide no information about the levels of uncertainty involved. However, to date, only a limited number of studies have evaluated interval forecasts in the tourism demand context. No study has evaluated and compared the density forecasts generated by alternative models of tourism demand. This study introduces the use of scoring rules for the evaluation of density forecasts and for comparisons between point and density forecasts.

The continuous rank probability score (CRPS) enables a direct comparison between point and density forecasts. The results of this scoring system indicate that density forecasts significantly outperform point forecasts. The SARIMA model is found to perform best. The superiority of this model is particularly evident for short-term forecasts. Although previous studies have shown that ISS and BSM models can provide acceptable interval forecasts, our study indicates that ISS and BSM models produce the worst density forecasts. These models are significantly outperformed by the SARIMA model over all forecast horizons. The AR and BOOT-AR models perform reasonably well, only being significantly outperformed by the SARIMA for short-term forecasts. In addition, bootstrapping is found to improve the accuracy of density forecasts, but only over short time horizons.

Density forecasts provide a complete probabilistic description of possible future realisations, thus providing more information than point or interval forecasts. In many cases decision makers have differing loss functions, and under such circumstances, point or interval forecasts no longer suffice. Hotel managers and government officials often need to know the probability that a specific event will occur in the future (for example, they may need to know the likelihood that the growth rate in tourist arrivals will exceed a certain number by a certain time). Such predictions are necessary for timely business and investment decisions. The comparison between point and density forecasts provides further evidence for the superiority of density forecasts. In addition, the rankings of alternative density forecasts, as determined in this study, suggest the use of the SARIMA model. These findings can help users to make better-informed decisions and choose the most appropriate models for their needs.

For most of the cases in this study, the absolute tests produce uniformly distributed PITs for all models, indicating that the assumption of normality for the density forecasts (except for bootstrapped versions) seems to be appropriate. However, in future studies using different datasets or models, it would be beneficial to test and select the distributional assumption of the density forecast first using absolute tests before doing the comparisons.

In contrast to the superior performance of BOOT-AR and poor performance of SARIMA in generating interval forecasts reported by Kim et al. (2011), this study finds that BOOT-AR only outperforms AR over short forecasting horizons and the SARIMA model provides the most accurate tourism demand density forecasts. Studies testing more origin–destination pairs are needed to determine whether these findings are generalisable. In addition, this study only uses time-series models. Causal models should be included as alternative models in future studies.

Funding

The financial support of National Natural Science Foundation of China (Grant No. NSFC71673233) is acknowledged. The second author would like to acknowledge the financial support from the International Doctoral Innovation Centre, Ningbo Education Bureau, Ningbo Science and Technology Bureau, and the University of Nottingham.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.annals.2018.12.019.

References

Alkema, L., Raftery, A. E., & Clark, S. J. (2007). Probabilistic projections of HIV prevalence using Bayesian melding. *The Annals of Applied Statistics, 1*(1), 229–248.
Athanasopoulos, G., & Hyndman, R. J. (2008). Modelling and forecasting Australian domestic tourism. *Tourism Management, 29*(1), 19–31.
Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting, 27*(3), 822–844.
Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics, 19*(4), 465–474.
Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control.* San Francisco: Holden-Day.
Croushore, D. D. (1993). *Introducing: The survey of professional forecasters.* Business Review-Federal Reserve Bank of Philadelphia November/December, 3–15.
Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review, 39*(4), 863–883.
Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics, 13*(3), 253–265.
Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69*(2), 243–268.
Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application, 1*, 125–151.
Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359–378.
Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological),* 107–114.
Greenidge, K. (2001). Forecasting tourism demand: An STM approach. *Annals of Tourism Research, 28*(1), 98–112.
Hammond, G. (2012). *State of the art of inflation targeting. Centre for central banking studies handbook No. 29.* Bank of England.
Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter.* Cambridge: Cambridge University Press.
Hong, Y., & White, H. (2005). Asymptotic distribution theory for nonparametric entropy measures of serial dependence. *Econometrica, 73*(3), 837–901.
Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 27*(3), 1–22.
Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: The state space approach.* Berlin: Springer Science and Business Media.
Jones, H. E., & Spiegelhalter, D. J. (2012). Improved probabilistic prediction of healthcare performance indicators using bidirectional smoothing models. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 175*(3), 729–747.
Kim, J. H. (2001). Bootstrap-after-bootstrap prediction intervals for autoregressive models. *Journal of Business and Economic Statistics, 19*(1), 117–128.
Kim, J. H. (2004). Bootstrap prediction intervals for autoregression using asymptotically mean-unbiased estimators. *International Journal of Forecasting, 20*(1), 85–97.
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.
Krüger, F., Lerch, S., Thorarinsdottir, T. L., & Gneiting, T. (2016). Probabilistic forecasting and comparative model assessment based on Markov chain Monte Carlo output. Working Paper. Available from: https://arxiv.org/abs/1608.06802.
Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics, 54*(1), 159–178.
Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics, 227*(7), 3515–3539.
Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science, 22*(10), 1087–1096.
Montgomery, J. M., Hollenbach, F. M., & Ward, M. D. (2012). Ensemble predictions of the 2012 US presidential election. *PS: Political Science and Politics, 45*(4), 651–654.
Osborn, D. R., Chui, A. P., Smith, J. P., & Birchenhall, C. R. (1988). Seasonality and the order of integration for consumption. *Oxford Bulletin of Economics and Statistics, 50*(4), 361–377.
Palmer, T. N. (2012). Towards the probabilistic Earth-system simulator: A vision for the future of climate and weather prediction. *Quarterly Journal of the Royal Meteorological Society, 138*(665), 841–861.
Peng, B., Song, H., & Crouch, G. I. (2014). A meta-analysis of international tourism demand forecasting and implications for practice. *Tourism Management, 45*, 181–193.
Raftery, A. E., Alkema, L., & Gerland, P. (2014). Bayesian population projections for the United Nations. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics, 29*(1), 58–68.

Shaman, P., & Stine, R. A. (1988). The bias of autoregressive coefficient estimators. *Journal of the American Statistical Association, 83*(403), 842–848.

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. (2000). *Tourism demand modelling and forecasting: Modern econometric approaches.* Oxford: Pergamon.

Tay, A. S., & Wallis, K. F. (2000). Density forecasting: A survey. *Journal of Forecasting, 19*(4), 235–254.

Vu, J. C., & Turner, L. W. (2006). Regional data forecasting accuracy: The case of Thailand. *Journal of Travel Research, 45*(2), 186–193.

Wan, S. K., Song, H., & Ko, D. (2016). Density forecasting for tourism demand. *Annals of Tourism Research, 60*, 27–30.

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.

Haiyan Song, PhD, is Chan Chak Fu Professor in International Tourism in the School of Hotel and Tourism Management at The Hong Kong Polytechnic University. His research interests are in the areas of tourism demand modeling and forecasting, tourism supply chain management and tourist satisfaction index research.

Long Wen, is a PhD candidate in the School of Economics, University of Nottingham Ningbo China. His research interests are mainly in tourism demand modelling and forecasting using various techniques including time series, econometric models and artificial neural networks.

Chang Liu, PhD, is associate professor and deputy head of School of Economics at University of Nottingham Ningbo China. His research interests include international trade, foreign direct investment and regional economics.