1. Introduction

As part of the vigorous development of the service industry, tourism has brought significant employment and income opportunities to destination economies. According to the World Travel & Tourism Council (WTTC), the total revenue of the global tourism industry exceeded US$8.9 trillion and generated 330 million tourism-related jobs in 2019 (WTTC, 2019). Accurate forecasts can provide useful information on future tourism demand and help practitioners, policymakers, and stockholders in the decision-making process (Song et al., 2012).

The general-to-specific (GETS) approach combined with the autoregressive distributed lag (ADL-GETS) specification is frequently used in tourism forecasting. The ADL-GETS model performs relatively well in forecasting tourism demand (Campos et al., 2005; Li et al., 2005; Narayan, 2004; Song & Witt, 2003). However, it suffers from the unstable decision rule problem. Small changes can significantly influence the model reduction process of GETS during the data training step, resulting in the possible exclusion of important explanatory variables from the final model.

Bagging, introduced by Breiman (1996), may offer a solution to the unstable rule problem. Bagging is especially useful for procedures such as decision tree models and artificial neural network models that are sensitive to small changes in training data. Bühlmann and Yu (2002) provided theoretical proof and empirical validation of the effectiveness of bagging for modelling unstable procedures. In recent years, researchers have verified that bagging (hereafter ordinary bagging) can overcome the unstable problem in the ADL-GETS model in tourism forecasting (Athanasopoulos et al., 2018).

However, ordinary bagging may not be effective when the sample size is small (Clyde & Lee, 2001). If the observations in the bootstrap sample do not contain enough variability, the models produced by the ordinary bagging procedure may differ from each other. Therefore, it can be difficult for ordinary bagging to estimate the true values (Fushiki, 2010).

The Bayesian bootstrap method was first proposed by Rubin (1981), and it has since proved to be more theoretically complete and effective than ordinary bagging, especially with small samples (Clyde & Lee, 2001; Lee & Clyde, 2004). In tourism forecasting, historical tourism demand series tend to be relatively short and highly volatile. To the best of our knowledge, no studies have used BBagging to forecast tourism demand. This paper verifies the performance of BBagging in forecasting tourism demand.

This paper is organised as follows. The next section reviews research on tourism demand modelling and forecasting with a particular focus on the GETS model, ordinary bagging, and BBagging. The third section describes the methodology and data. The fourth section presents the empirical results, and the last section offers conclusions and research limitations.

2. Literature Review

Over the last two decades, a variety of methods have been used to forecast tourism demand (Song et al., 2019). Tourism forecasting models can be divided into three categories: time series, econometric, or artificial intelligence (Wu et al., 2017). Econometric models have an advantage over time series models in that they can postulate causal relationships between tourism demand and explanatory variables. Econometric models also allow for the estimation of tourism demand elasticities, which can provide useful information for policymakers. In forecasts of destination-level tourism demand, the main influencing factors are the income (or per capita income) of the origin country/region, the price of the destination relative to the price of the origin country/region, and the prices of the substitute destinations (Crouch, 1992; Song & Li, 2008). These factors are derived from neoclassical demand theory in economics and therefore have a strong theoretical foundation. Peng et al. (2014) showed that, in general, econometric models are superior to the other two types of models in terms of forecasting performance.

2.1 The ADL-GETS model

A number of studies have applied GETS (Hendry, 1995) to tourism demand forecasting (Song & Witt, 2003). GETS modelling is an extension of the general tourism demand model that includes all of the potential explanatory variables suggested by demand theory, the lagged dependent and explanatory variables that account for the dynamic features of the time series, and one-off event dummies. This model specification is known as the ADL model and is estimated by the ordinary least squares (OLS) method. The general ADL model is recursively estimated to eliminate non-significant or incorrectly signed variables until it reaches the final specification, in which all of the variables play an important role in determining the dependent variable (Song, Witt, & Li, 2003).

The GETS approach, particularly the ADL-GETS model, has been widely used in tourism demand modelling and forecasting for different destinations, such as Hong Kong (Song, Wong & Chon, 2003), Thailand (Song, Witt & Li, 2003), Fiji (Narayan, 2004), and China (Lin et al., 2015; Song & Fei, 2007). The GETS model has also been used to investigate the influence of the 2008 Global Financial Crisis on the tourism and hospitality industry in Asia (Song et al., 2010; Song et al., 2011; Song & Lin, 2010) and the UK (Page et al., 2012), and to examine the global impact of terrorism attacks on inbound arrivals (Liu & Pratt, 2017). Liu et al. (2020) summarised the performance of the ADL-GETS model in a large-scale ex ante forecasting project. They found that the variation in the tourism demand and gross domestic product (GDP) of source markets and the covariation between tourism demand and GDP have significant effects on the forecasting accuracy of the ADL-GETS model over different horizons. In other words, some factors, such as

fluctuations in historical data, can affect the stability of the forecasting performance in ways that are beyond the control of the model.

A number of tourism forecasting researchers are trying to address the forecasting performance instability of the ADL-GETS model. Wong et al. (2007) and Song, Witt, Wong, and Wu (2009) combined the forecasts of the ADL-GETS model with other methods to improve forecasting stability. Li et al. (2019) combined interval forecasts to enhance the stability. Lin et al. (2014) introduced judgemental forecasts to integrate experts' opinions into the statistical forecasts generated by the ADL-GETS model. The latter approach has been adopted to establish a web-based tourism demand forecasting system that integrates a database, an ADL modelling procedure, and a judgmental adjustment module (Song et al., 2008, 2013). Although the above methods use advanced econometric techniques to improve the forecasting performance of the single ADL-GETS model, the forecasts are still generated based on the one-off estimation of the ADL-GETS model. Thus, these methods do not address the instability of the ADL-GETS model.

2.2 *Ordinary Bagging*

Breiman (1996) and Bühlmann and Yu (2002) proposed using the ordinary bagging method to improve the instability of the GETS model. Ordinary bagging involves generating several distinct training series from the original dataset, estimating a defined model multiple times, and producing forecasts based on the estimated models. The final forecasts are calculated by averaging all of the forecasts generated by the estimated models. Bühlmann and Yu (2002) provided a theoretical illustration of how ordinary bagging could improve prediction stability and reduce variance in regression-based predictions in difficult decision problems. Inoue and Kilian (2008) elaborated on how ordinary bagging can reduce errors in out-of-sample inflation forecasts using the GETS algorithm. Lee and Yang (2006) also demonstrated that combining ordinary bagging with asymmetric cost functions can improve binary prediction accuracy, which they verified with Monte Carlo simulations and an empirical study of S&P500 and NASDAQ stock indices. In addition, Petropoulos et al. (2018) showed that ordinary bagging reduces uncertainties associated with the data generating process, model specification, and model parameter estimation in time series forecasting.

Recent tourism studies have introduced ordinary bagging to the ADL-GETS modelling and forecasting process. For instance, Athanasopoulos et al. (2018) demonstrated that ordinary bagging improves the ADL-GETS model's performance in Australian tourism demand forecasting. However, Song et al. (2017) found that ordinary bagging fails to realise the expected improvement when it is incorporated into the ADL-GETS model using quarterly data on Hong Kong visitor arrivals as sample sets. This contradiction reveals the potential problems caused by high volatility in tourism datasets at the macro and micro levels. They are sensitive to

represents the ADL-GETS model.

3. The final forecast is $\hat{y}^*_{n+h} = median\left(\hat{y}^*_{b,n+h}\right)$ or $mean(\hat{y}^*_{b,n+h})$.

The resampling progress can also be seen as assigning weight ω_i to each bootstrap sample, $\omega \in \{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\}$. The sum of ω_i is equal to 1, and $n\omega_i$ is the frequency with which M_i appears in the bootstrap sample. Thus, when ordinary bagging is applied to a small sample (n), ω may become more diverse and induce greater variance in the prediction.

As all of the tourism demand datasets are time series that include trends and seasonality, classic bagging (bootstrapping a sample with replacement) cannot retain all of the information from the original series. In our study, the function 'bld.mbb.bootstrap' in R is used to generate bootstrap samples, as proposed by Bergmeir et al. (2016). In this method, the bootstrap procedure is only implemented on the residual term, which is regarded as the white noise generated by STL decomposition. Thus, the trend and seasonality of the original data are preserved. Outliers generated by one-off events are smoothed before the bootstrap procedure is implemented, which ensures that the effects of one-off events do not randomly appear in the bootstrap series and thus affect the bagging result.

3.3 BBagging

BBagging treats the weight ω_i of each bootstrap sample as unknown and follows the non-informative Dirichlet prior, $\pi(\omega) \propto \prod_{i=1}^{n} \omega_i^{-1}$ (Rubin, 1981). The posterior distribution $\pi(\omega|M_n)$ follows a Dirichlet distribution when the prior is combined with a multinominal likelihood. The Dirichlet distribution is the conjugate prior of the multinomial distribution, where all of the values are between 0 and 1, and their sum is 1. An n-dimensional Dirichlet distribution is represented by $Dirichlet_n(\alpha_1, \alpha_2, \dots, \alpha_n)$, where the expectation of a dimension equals $\alpha_i / \sum \alpha_{1,\dots,n}$. The distribution becomes closer to the proportion of the parameter mean as $\sum \alpha_{1,\dots,n}$ increases (Newton & Raftery, 1994). In the non-informative prior, the weights of all of the data are expected to be equal, which indicates that uniform Dirichlet weights should be used.

For a model that can accept weight, we can obtain a predicted value of $\hat{y}^*_{bb,t+H} = f(M_n, \omega_{b,n})$ for each set of $\omega_{b,n}$ generated from $\pi(\omega|M_n)$.

The ADL model can be written as

$$\underset{(1\times o)}{\boldsymbol{Y}} = \underset{(1\times o)}{\boldsymbol{X}} \cdot \underset{(o\times 1)}{\boldsymbol{\eta}} + \boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{X_i} = [1, y_{n-1}, \dots, y_{n-p}, x_{k,n}, \dots, x_{k,n-q}, D_1, \dots, D_m]$,$\boldsymbol{\eta} = [\alpha, \beta_1, \dots \beta_p, \gamma_0, \dots, \gamma_q, \varphi_1, \dots \varphi_m]'$.

Then, the OLS method is used to estimate parameter $\widehat{\boldsymbol{\eta}}$:

$$\widehat{\boldsymbol{\eta}} = \arg\min\|\boldsymbol{Y} - \boldsymbol{X\eta}\|^2 = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{3}$$

In practice, the BBagging procedure is the same as the weighted least squares estimation (MLS), which is as follows:

$$\widehat{\boldsymbol{\eta}}_b^* = \arg\min(\boldsymbol{Y} - \boldsymbol{X\eta})^T\boldsymbol{W}_b(\boldsymbol{Y} - \boldsymbol{X\eta}) = (\boldsymbol{X}'\boldsymbol{W}_b\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}_b\boldsymbol{Y} \tag{4}$$

$\boldsymbol{W}_b = \begin{bmatrix} \omega_{b,1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_{b,n} \end{bmatrix}$ is a diagonal matrix constructed by $\omega_{b,i}$. When all of the $\omega_{b,i}$ equal $\frac{1}{n}$, the estimates are equivalent to OLS. The posterior distributions of $\widehat{\boldsymbol{\eta}}$ and $\hat{y}_{bb,t+H}^*$ are easily obtained by running the Monte Carlo estimation B times. The final forecast is the median or mean of the predictive distribution (Clyde & Lee, 2001).

3.4 Data and modelling process

To eliminate the impact of policy interventions on visitor arrivals from the prediction, this study uses the annual and quarterly visitor arrivals to Hong Kong from six visa-free origin countries (Australia, Canada, France, Germany, the UK, and the US). The visitor arrivals data are obtained from the Hong Kong Tourism Board's PartnerNet (https://partnernet.hktb.com/en/home/index.html). To compare the performance of BBagging in different samples, the annual and quarterly data are treated as small and large samples, respectively. Based on demand theory, in addition to the income of the origin country, the relative price of tourism products in Hong Kong and the origin country (adjusted by exchange rates; Song, Witt, & Li, 2003) is included in the model to capture the effect of prices on tourism demand. The income of the origin country is measured by the real GDP of the source markets, and the relative price variable is calculated as follows:

$$RP_{e,i} = \frac{CPI_{HK,i}/EX_{HK,i}}{CPI_{e,i}/EX_{e,i}} \tag{5}$$

where $CPI_{HK,i}$ and $CPI_{e,i}$ are the consumer price indices (CPIs) of Hong Kong and the source country e, respectively. $EX_{HK,i}$ and $EX_{e,i}$ represent the exchange rate of the Hong Kong dollar against the currency of the source country measured in US dollars. GDP, CPI, and exchange rate (EX) all use the 2010 values as a base. These data are collected from the *International Financial Statistics* of the International Monetary Fund (IMF). Seasonal dummy variables and one-off events, such as the September 11, 2001 attacks, severe acute respiratory syndrome (SARS), and the 2008 Global Financial Crisis, are included in the models. Following tourism demand theory (Li et al., 2005), visitor arrivals, income, and relative price are log-transformed and represented as *LAR*, *LY*, and *LRP*, respectively.

To evaluate the performance of BBagging in samples with different lengths, annual and quarterly forecasts are conducted within different rolling windows. For annual forecasting, 1999 to 2014 is used for model estimation, and one- to four-step-ahead rolling forecasts are generated for the 2015–2018 period. For quarterly forecasting, the sample period is the same as for annual forecasting. The subsample from 1999Q1 to 2015Q4 is used for model estimation, and the remainder of the sample is used to evaluate the performance of the 1-step-ahead to 12-steps-ahead forecasts.

3.5 Forecasting evaluation

The evaluation of forecasting performance is based on *ex post* forecasts. The forecasting evaluation metrics are the mean absolute percentage error (MAPE) and the root mean square error (RMSE), which are commonly used in the literature (Wu et al., 2017). MAPE and RMSE diagnose the deviation of a prediction from the absolute and quadratic levels, respectively. These forecasting error measures are defined as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|e_i|}{y_i} \tag{6}$$

where $e_i = y_i - \hat{y}_i$, $\hat{y}_i$ is the forecast value of y_i, and

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}{e_i}^2} \tag{7}$$

4. Findings and Discussion

4.1 Unit root and cointegration tests

Before modelling and forecasting, unit root and bounds tests are conducted. Tables 1 and 2 present the integrated orders of all of the dependent and independent variables from the Augmented Dickey–Fuller (ADF) test, Phillips–Perron (PP) test, and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests. Tables 1 and 2 indicate that almost all of the time series are either *I (0)* or *I (1)* and thus eligible for the bounds test. The null hypothesis (H_0) of the bounds test is that there is no cointegration between the dependent and independent variables. If the *F*-statistic is greater than the given upper bound *I (1)* value, H_0 is rejected. If it falls between the bounds of *I (0)* and *I (1), t*-statistics are needed to confirm the finding. Details of the bounds test can be found in Pesaran et al. (2001) and Song and Lin (2010). Table 3 shows the results of the ADL model bounds test and the lower and upper bound values at the 1%, 5%, and 10% significance levels, obtained from Pesaran et al. (2001). As there are two independent variables, income and relative price, in the models that use annual data, the population parameter *k* equals 2. There are five independent variables in the models based on

quarterly data, as three additional seasonal dummy variables are introduced to capture the seasonality effect; k is therefore set to 5. The F-statistics in Table 3 indicate that the annual models for all of the origin countries show a significant long-term relationship between the dependent and independent variables at the 1% significance level, except for the samples from Canada and the UK, which are significant at the 10% level. The cointegration relationships are observed in all six source markets in the quarterly models at the 1% significance level. To verify the specifications of the ADL models used in the bounds tests, several diagnostic tests including the Durbin–Watson autocorrelation test (DW test), White's heteroskedasticity test (White test), ARCH heteroskedasticity test (Arch test), Jarque–Bera normality test (JB test), and Ramsey misspecification test (RESET test) are conducted. All of the models pass the diagnostic tests at the 5% significance level with the exceptions of the DW and JB tests in the quarterly US model and the RESET test in the annual Germany and quarterly Australia models. Thus, overall, the estimation results of the ADL models are valid, and the bounds test results are reliable. Due to space constraints, the diagnostic test results are omitted, but are available upon request.

Table 1. Unit Root Test Results for Annual Data

			Australia			
	LAR	ΔLAR	LY	ΔLY	LRP	ΔLRP
ADF	-1.076	-3.276***	-1.687	-2.602**	-1.421	-2.974***
PP	-7.242	-25.336***	-2.738	-18.231**	-4.597	-13.342
KPS	0.722**	0.141	0.892***	0.300	0.548**	0.156
Order	-	*I (1)*	-	*I (1)*	-	*I (1)*
			Canada			
	LAR	ΔLAR	LY	ΔLY	LRP	ΔLRP
ADF	-1.705	-3.518***	-1.911*	-	-1.725	-2.372**
PP	-8.327	-27.470***	-4.663	-22.015**	-2.191	-10.526
KPS	0.771***	0.177	0.877***	0.269	0.430*	0.234
Order	-	*I (1)*	-	*I (1)*	-	*I (1)*
			France			
	LAR	ΔLAR	LY	ΔLY	LRP	ΔLRP
ADF	-1.253	-3.676***	-2.289**	-	-2.174**	-
PP	-9.202	-26.275***	-3.608	-14.907	-5.623	-10.322
KPS	0.664**	0.118	0.837***	0.290	0.162	-
Order	-	*I (1)*	-	*I (1)*	*I (0)*	-
			Germany			
	LAR	ΔLAR	LY	ΔLY	LRP	ΔLRP
ADF	-3.068***	-	0.751	-3.950***	-2.164**	-
PP	-7.874	-24.948***	-2.665	-18.051**	-5.809	-10.505
KPS	0.138	-	0.812***	0.324	0.162	-
Order	*I (0)*	-	-	*I (1)*	*I (0)*	-
			UK			
	LAR	ΔLAR	LY	ΔLY	LRP	ΔLRP
ADF	-0.865	-3.271***	-1.443	-1.987*	-1.897*	-
PP	-11.833	-28.879***	-4.365	-13.612	-2.222	-14.221
KPS	0.677**	0.065	0.816***	0.251	0.211	-
Order	-	*I (1)*	-	*I (1)*	*I (0)*	-
			US			
	LAR	ΔLAR	LY	ΔLY	LRP	ΔLRP
ADF	-1.535	-4.630***	-1.412	-2.500**	-3.226***	-
PP	-20.195**	-	-5.443	-12.926	-1.791	-7.591
KPS	0.790***	0.075	0.874***	0.308	0.411*	0.282
Order	-	*I (1)*	-	*I (1)*	-	*I (1)*

Note. *, **, and *** represent the rejection of H_0 at the 10%, 5%, and 1% significance levels, respectively.

Table 2. Unit Root Test Results for Quarterly Data

			Australia			
	LAR	ΔLAR	LY	ΔLY	LRP	ΔLRP
ADF	-2.478**	-	-1.031	-8.865***	-1.779*	-
PP	-33.105***	-	-49.505***	-	-2.682	-53.142***
KPS	1.318***	0.054	2.058***	0.200	1.081***	0.353
Order	*I (0)*	-	-	*I (1)*	-	*I (1)*
			Canada			
	LAR	ΔLAR	LY	ΔLY	LRP	ΔLRP
ADF	-3.724***	-	-1.837*	-	-1.753*	-
PP	-48.774***	-	-39.383***	-	-2.298	-54.035***
KPS	1.308***	0.047	2.031***	0.215	0.672**	0.638**
Order	*I (0)*	-	*I (0)*	-	-	*I (1)*
			France			
	LAR	ΔLAR	LY	ΔLY	LRP	ΔLRP
ADF	-3.195***	-	-1.721*	-	-1.613	-5.613***
PP	-47.690***	-	-46.546***	-	-2.554	-52.152***
KPS	1.167***	0.049	1.889***	0.239	0.500**	0.317
Order	*I (0)*	-	*I (0)*	-	-	*I (1)*
			Germany			
	LAR	ΔLAR	LY	ΔLY	LRP	ΔLRP
ADF	-4.472***	-	-0.918	-14.098***	-1.658	-5.601***
PP	-61.035***	-	-13.199	-54.233***	-2.717	-52.738***
KPS	0.804***	0.042	1.766***	0.148	0.504**	0.295
Order	*I (0)*	-	-	*I (1)*	-	*I (1)*
			UK			
	LAR	ΔLAR	LY	ΔLY	LRP	ΔLRP
ADF	-4.358***	-	-1.939*	-	-1.492	-7.057***
PP	-51.135***	-	-5.861	-84.102***	-3.960	-45.770***
KPS	1.258***	0.043	1.681***	0.244	0.553**	0.341
Order	*I (0)*	-	-	*I (1)*	-	*I (1)*
			US			
	LAR	ΔLAR	LY	ΔLY	LRP	ΔLRP
ADF	-4.125***	-	-0.844	-4.152***	-3.007***	-
PP	-68.360***	-	-7.212	-54.988***	-4.178	-69.954***
KPS	0.961***	0.030	2.005***	0.147	0.497**	1.332***
Order	*I (0)*	-	-	*I (1)*	-	-

Note. *, **, and *** represent the rejection of H_0 at the 10%, 5%, and 1% significance levels, respectively.

Table 3. Bounds Test Results of Annual and Quarterly Models

	Annual Models		Quarterly Models	
Country	*F*-Statistic	*t*-Statistic	*F*-Statistic	*t*-Statistic
Australia	16.56***	-6.19***	5.13***	-3.96*

Canada	4.76*		-3.67**		19.99***		-5.44***	
France	13.06***		-6.18***		8.18***		-5.66***	
Germany	11.15***		-5.31***		7.29***		-4.81***	
UK	4.76*		-3.39*		16.67***		-5.92***	
US	7.15***		-4.62***		11.74***		-6.87***	
Critical Value of Bounds Test	*I (0)*	*I (1)*	*I (0)*	*I (1)*	*I (0)*	*I (1)*	*I (0)*	*I (1)*
10% Significance Level	3.17	4.14	-2.57	-3.21	2.26	3.35	-2.57	-3.86
5% Significance Level	3.79	4.85	-2.86	-3.53	2.62	3.79	-2.86	-4.19
1% Significance Level	5.15	6.36	-3.43	-4.10	3.41	4.68	-3.43	-4.79

Note. *, **, and *** represent the rejection of H_0 at the 10%, 5%, and 1% significance levels, respectively.

4.2 Estimation results

Tourism demand elasticity is a measure of how tourism demand responds to changes in the independent variables. Based on demand theory, income elasticity is expected to be positive, whilst price elasticity is expected to be negative. Table 4 shows the long-term income and price elasticities in the annual and quarterly models for all of the source countries. Some of the elasticities are not available in the annual models because the non-significant variables (*LY* or *LRP*) are removed by the GETS procedure. The correct signs of income and price elasticities indicate that the ADL-GETS models are able to capture the long-term relationships between variables at different data frequencies.

Table 4. Tourism Demand Elasticities in Annual and Quarterly Models

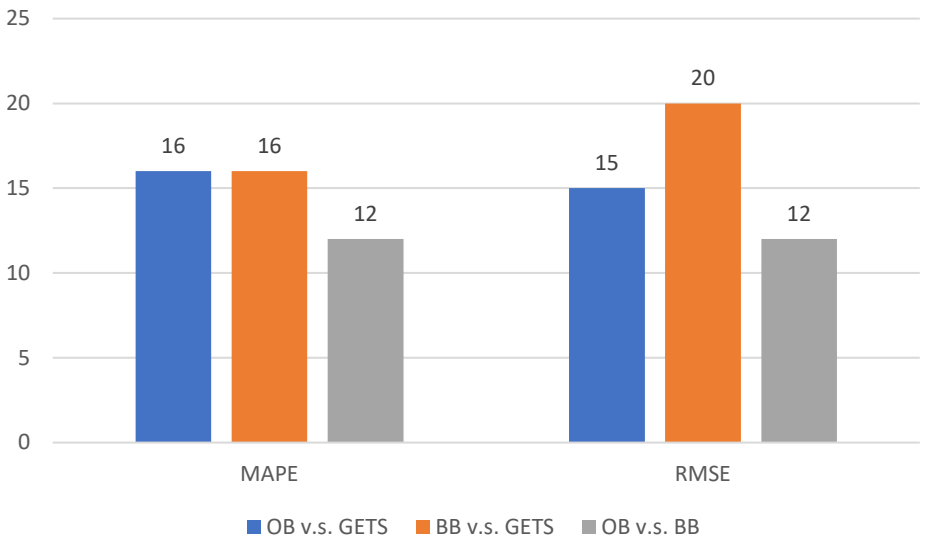
	Annual		Quarter	
Country	Income Elasticity	Price Elasticity	Income Elasticity	Price Elasticity
Australia	1.62	-	0.04	-0.82
Canada	-	-0.85	0.53	-0.60
France	1.30	-	1.54	-0.69
Germany	-	-0.48	0.39	-0.41
UK	1.00	-0.56	1.73	-0.25
US	0.42	-0.82	0.42	-0.64

4.3 Annual forecasts

The annual forecasts of the visitor arrivals from the six long-haul source markets to Hong Kong are implemented using ADL-GETS, GETS Bagging, and GETS BBagging. The forecasting performance of the three methods, measured by MAPE and RMSE, are presented in Table 5. The grid search suggests that a Dirichlet

distribution in which α_i equals 7 produces the most accurate forecasts based on both MAPE and RMSE when the hyperparametric is set to search between 1 and 10. The models use different procedures to predict the tourism demand of the six markets for the 1-year-ahead to 4-years-ahead forecast horizons. In general, the forecasting accuracy of all of the models decreases as the forecast horizon extends, due to increasing uncertainty. For example, the MAPE values of the GETS model for Germany increase from 0.89% to 12.15% when the forecast horizon extends from 1-year ahead to 4-years ahead. This finding is consistent with the tourism forecasting literature (Song et al., 2010; Song, Wong & Chon, 2003). For individual source markets, the three methods produce different results. For example, in the cases of Australia, Canada, and the UK, the Bagging models outperform the simple GETS model across all horizons according to both error measures. In the cases of France and Germany, ordinary bagging fails to improve GETS forecasting accuracy, but BBagging shows significant improvements except in the 1-year-ahead forecasts. For the U.S. case, when the performance is measured by MAPE, ordinary bagging beats GETS in all horizons, with the exception of the 1-year-ahead forecast, but BBagging does not improve the performance of GETS in any of the four horizons. However, when RMSE is adopted, there is no clear winner among the three competing models.

To verify whether the Bagging methods can improve the GETS model's performance, the two evaluation measures are calculated by the median of the bagging prediction distribution (see Figure 1 and Table 5). The results suggest that, overall, both bagging methods improve the forecasting accuracy of the single GETS model. When the forecasts are evaluated by relative measurements (i.e., MAPE), the performance of ordinary bagging and BBagging is similar. Across all six markets and four forecast horizons (24 cases), both bagging methods reduce the forecast errors of the GETS model in 16 out of 24 cases according to the MAPE values. According to the one-sample Wilcoxon test, the Wilcoxon statistic for both ordinary bagging and BBagging is 136 and significant at the 0.1% significance level, which suggests that the improvement is statistically significant. When forecasting accuracy is measured by the absolute error (i.e., RMSE), ordinary bagging can improve GETS' forecasting performance in 15 out of 24 cases, whilst BBagging can improve GETS' forecasting performance in 20 out of 24 cases, with Wilcoxon statistics of 120 ($p = 0.000$) and 210 ($p = 0.000$), respectively. Thus, overall, both ordinary Bagging and BBagging can improve the forecasting accuracy of GETS in annual forecasting.



Note. OB = ordinary bagging; BB = BBagging.

Figure 1. Performance of the Three Methods in Annual Forecasting

Table 5. Annual Forecasting Performance of the GETS, Ordinary Bagging, and BBagging Models

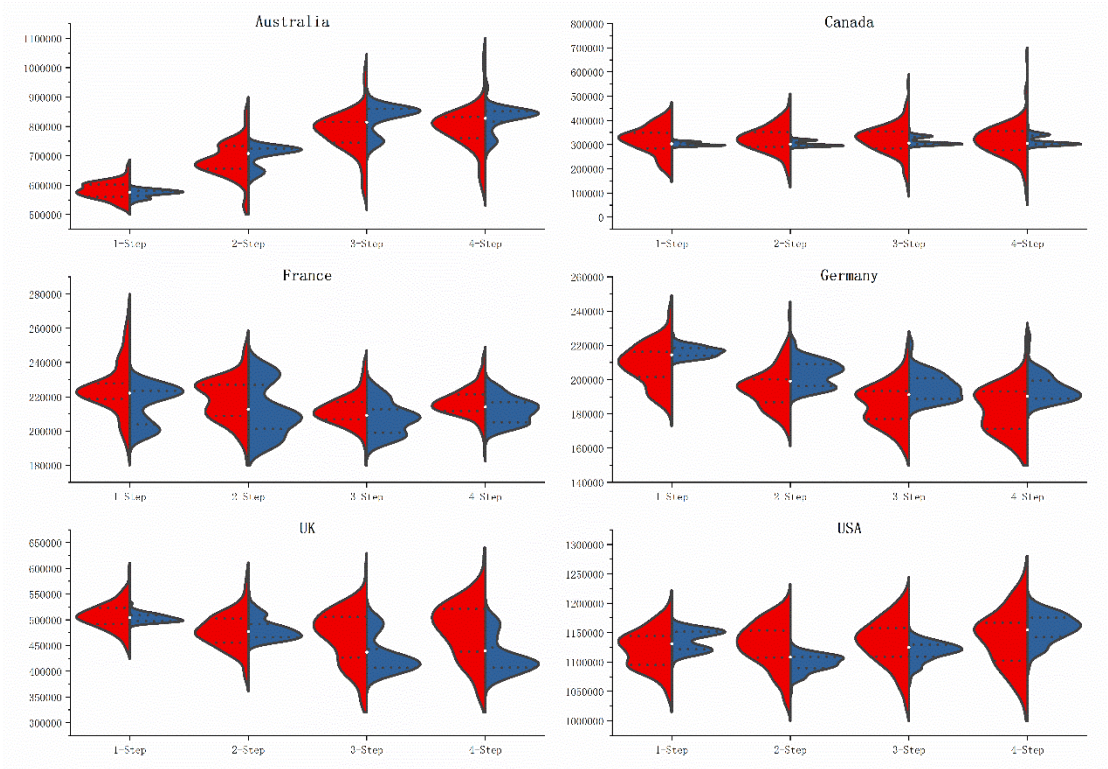
Country	Horizon	MAPE			RMSE		
		GETS	GETS.OB	GETS.BB	GETS	GETS.OB	GETS.BB
Australia							
	1	0.75	1.16	*0.03*	4.35	6.74	*0.17*
	2	10.59	6.63	*5.39*	104.58	40.83	*32.79*
	3	18.23	*15.04*	20.08	186.01	*101.41*	143.88
	4	21.51	*20.48*	21.68	208.57	*148.01*	159.06
Canada							
	1	20.44	*10.73*	19.93	60.83	*34.75*	59.56
	2	22.48	*12.74*	21.92	67.05	*41.48*	65.65
	3	22.34	*12.97*	21.21	67.00	*42.37*	64.27
	4	22.93	*14.29*	21.94	69.05	*46.67*	66.76
France							
	1	*4.36*	6.08	4.83	*8.77*	13.58	10.65
	2	6.62	4.83	*0.90*	13.78	10.92	*1.92*
	3	5.35	3.90	*1.58*	11.71	9.34	*3.91*
	4	4.56	4.62	*2.16*	10.39	11.07	*5.42*
Germany							
	1	*0.89*	2.09	1.25	*1.92*	4.37	2.70
	2	6.45	9.86	*3.73*	17.24	20.77	*10.18*
	3	9.92	13.50	*7.76*	23.44	27.00	*16.97*
	4	12.15	15.54	*10.50*	27.13	30.49	*21.89*
UK							
	1	5.69	*4.57*	5.34	28.50	*23.15*	26.82
	2	11.04	9.04	*8.92*	58.56	46.22	*45.66*
	3	18.44	*11.69*	12.55	93.17	*58.24*	61.90

	4	23.38	*13.20*	17.27	113.04	*66.21*	82.79
US							
	1	*2.94*	5.17	3.39	*33.77*	58.06	38.69
	2	6.60	*6.13*	7.33	83.18	*70.81*	83.10
	3	7.36	*6.36*	7.68	88.79	*73.57*	87.21
	4	8.58	*8.48*	9.31	*104.77*	106.39	114.25

Note. 1. The magnitude of MAPE is measured as a percentage, and the magnitude of all RMSE values is given at 10^6. 2. OB = ordinary bagging; BB = BBagging. 3. The numbers in italics indicate the best performing model in each horizon, as assessed by MAPE and RMSE.

A close examination of Figure 1 shows that ordinary bagging outperforms BBagging in 12 out of 24 cases across the two error measures. However, the one-sample Wilcoxon test shows no significant difference between the two methods (W = 137.5, p = 0.671), indicating that although BBagging can statistically beat GETS in annual forecasting, it does not statistically outperform ordinary bagging in its predictions of changes in Hong Kong's six long-haul markets.

Forecasting accuracy is a key indicator to assess forecasting performance, but the variation in the forecasts should not be ignored, particularly when the performance of GETS is not stable (Bühlmann & Yu, 2002). A split violin chart is used to further compare the variability in the prediction distribution of the two bagging methods (see Figure 2). The split violin chart maps the nuclear density of the 100 forecasts generated by the ordinary bagging and BBagging procedures for all six source markets. The lower and upper dotted lines represent the 25% and 75% quantiles, respectively. The white dot on the central axis represents the median of each forecasting distribution. The shape of the violin chart and the box widths show that BBagging generates a more centralised prediction distribution than ordinary bagging.



Note. The red chart is the ordinary bagging distribution and the blue chart is the BBagging distribution.

Figure 2. Distributions of Annual Forecasts Using Ordinary Bagging and BBagging

Table 6 presents the coefficients of variation of the forecast errors in the four forecast horizons across the six source markets generated by the two bagging methods. Consistent with the graphs in Figure 1, the Wilcoxon test suggests that the variation in the BBagging forecasts is significantly smaller than the variation in the ordinary bagging forecasts (W = 456, p = 0.000), indicating that BBagging generates more concentrated forecasts than ordinary bagging.

Table 6. Coefficients of Variation of Annual Forecast Errors

Horizon	Australia		Canada		France	
	OB	BB	OB	BB	OB	BB
1	0.049	0.022	0.175	0.025	0.066	0.051
2	0.088	0.050	0.168	0.037	0.057	0.070
3	0.092	0.056	0.196	0.050	0.044	0.044
4	0.095	0.052	0.234	0.061	0.040	0.035
Horizon	Germany		UK		US	
	OB	BB	OB	BB	OB	BB
1	0.057	0.015	0.051	0.016	0.029	0.015
2	0.060	0.036	0.073	0.042	0.034	0.013
3	0.066	0.040	0.111	0.080	0.035	0.014
4	0.075	0.040	0.117	0.080	0.043	0.020

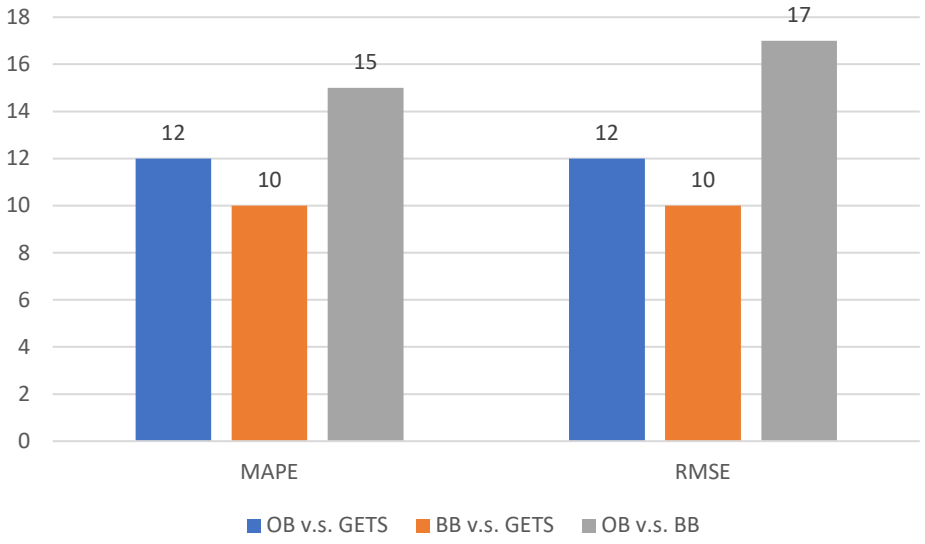
Abbreviation: BB, BBagging; OB, ordinary bagging.

Overall both bagging methods can statistically beat GETS and have similar forecasting performance in the annual forecasts. Although BBagging does not demonstrate a more accurate performance than ordinary bagging, as suggested by Kurogi and Harashima (2009), it reveals a more stable result with stronger validity, which is consistent with Fushiki (2010). Thus, BBagging can not only beat the simple GETS model but is also preferable to ordinary bagging due to its higher validity when using small samples for forecasting.

4.4 Quarterly forecasts

The grid search from 1 to 10 suggests that the quarterly forecasts are most accurate when the hyperparameter in the Dirichlet distribution equals 3. The summary of the forecasting performance is presented in Figure 3 and Table 7. Six selected quarterly forecast horizons are presented in Table 7, covering short-run forecasts (i.e., one-quarter to three-quarters ahead) and medium-run forecasts (i.e., 6- to 12-quarters ahead). Consistent with the results for the annual forecasts, in general, as the forecast horizon extends, the forecast errors become larger. However, in the quarterly forecasts, the two bagging methods only beat GETS in the case of France, in four out of six horizons. In the other five source markets, the bagging methods fail to outperform GETS.

In the 36 (6 horizons × 6 source markets) quarterly forecasts, ordinary bagging outperforms GETS 12 times when the performance is measured by MAPE and RMSE, whilst BBagging beats GETS 10 times across the two error measures. The one-sample Wilcoxon test suggests that neither ordinary bagging nor BBagging significantly improves the performance of GETS for the quarterly forecasts. It is not surprising that the performance of GETS tends to improve and become more stable as the sample size increases from 16 periods in the annual forecasts to 68 periods in the quarterly forecasts (Liu et al., 2020). Thus, the superiority of BBagging in handling high forecast variability in small samples is not fully reflected in the quarterly forecasts. Ordinary bagging in quarterly forecasting cannot beat GETS, which contradicts the findings of Athanasopoulos et al. (2018), perhaps because the sample size in this study is much smaller than theirs. Although GETS becomes more stable in larger samples (Liu et al., 2020), ordinary bagging can be much more accurate and beats GETS when the model is fitted with large samples. The advantages of GETS over ordinary bagging should be examined in future studies.



Note. OB = ordinary bagging; BB = BBagging.

Figure 3. Performance of the Three Methods in Quarterly Forecasting

Table 7. Quarterly Forecasting Performance of GETS, Ordinary Bagging, and BBagging

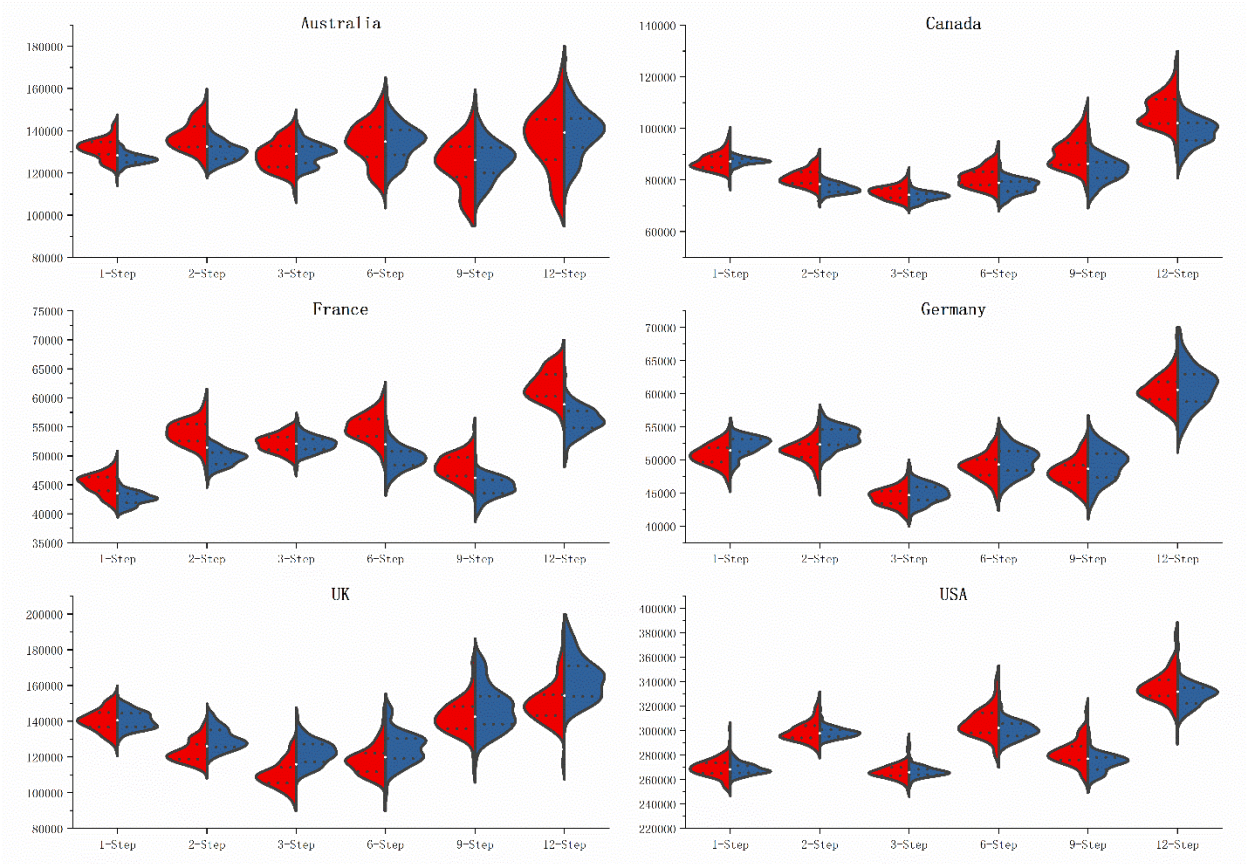
Country	Horizon	MAPE			RMSE		
		GETS	OB	BB	GETS	OB	BB
Australia							
	1	1.03	*0.77*	2.20	1.37	*1.04*	2.89
	2	*6.30*	7.33	8.00	11.39	*11.19*	14.83
	3	8.21	7.25	*6.40*	13.86	*10.79*	11.53
	6	*6.50*	8.30	8.48	11.91	*12.36*	12.45
	9	*7.93*	8.46	8.44	14.20	12.79	*12.57*
	12	9.03	8.19	*8.12*	16.32	13.46	*13.13*
Canada							
	1	10.17	9.71	*9.32*	8.86	8.50	*8.18*
	2	10.75	*8.00*	10.93	8.82	*8.28*	10.19
	3	*8.33*	9.77	12.17	*7.36*	9.07	11.73
	6	*10.48*	11.25	15.19	*9.66*	12.68	15.27
	9	*10.66*	13.66	16.60	*10.27*	14.34	16.63
	12	*9.72*	13.64	16.64	*9.65*	14.34	16.78
France							
	1	13.76	*7.49*	13.91	5.80	*3.34*	5.86
	2	11.47	*6.19*	8.88	5.21	*3.47*	5.16
	3	10.78	*6.44*	4.85	5.15	3.63	*2.82*
	6	10.22	8.72	*7.71*	*5.25*	5.43	5.87
	9	*8.06*	9.75	8.44	*4.56*	5.76	5.58
	12	*7.41*	10.06	8.94	*4.57*	6.24	5.34
Germany							
	1	*10.40*	14.98	10.93	*5.50*	7.61	5.76
	2	7.54	8.35	*5.15*	4.30	5.10	*3.40*

	3	*5.76*	9.85	7.94	*3.56*	5.51	4.88
	6	*7.06*	14.62	11.11	*4.28*	9.01	7.61
	9	*6.52*	16.16	12.72	*4.08*	9.54	8.10
	12	*6.42*	16.45	13.36	*4.04*	10.07	8.79
UK							
	1	5.14	*2.85*	2.87	7.08	*4.01*	4.03
	2	5.37	4.98	*4.82*	7.02	9.14	*6.83*
	3	*4.47*	9.37	6.90	*6.00*	13.80	10.59
	6	*7.62*	9.93	8.04	*10.95*	15.05	12.88
	9	*7.47*	11.28	9.96	*10.77*	17.09	15.48
	12	*7.19*	12.09	10.16	*10.30*	18.01	15.76
US							
	1	5.43	4.09	*4.08*	14.38	10.97	*10.94*
	2	*5.71*	5.82	5.80	*16.09*	20.68	20.28
	3	*4.23*	5.82	6.36	*13.28*	21.92	24.13
	6	*6.99*	9.02	9.08	*23.00*	32.92	33.16
	9	*6.70*	9.38	10.42	*22.67*	32.82	36.75
	12	*7.33*	10.40	11.04	*26.09*	38.06	40.50

Note. 1. The magnitude of MAPE is given as a percentage. The magnitude of all RMSE values is 10^6. 2. OB = ordinary bagging; BB = BBagging. 3. The numbers in italics indicate the best performing model in each horizon, assessed by MAPE and RMSE.

Although neither of the two bagging methods are able to beat GETS in quarterly forecasting, BBagging is more accurate than ordinary bagging in 21 out of 36 cases when measured by MAPE. The performance of BBagging and ordinary bagging is close when measured by RMSE, as the former only beats the latter in 17 out of 36 cases. The one-sample Wilcoxon test indicates that the improvement is not significant at the 5% significance level according to either error measure. The comparison of the two methods in quarterly forecasting suggests that ordinary bagging and BBagging have similar forecasting performance, which is consistent with the finding of the annual forecasts.

Figure 4 and Table 8 present the variations in the quarterly forecasts generated by ordinary bagging and BBagging. The violin plots in Figure 4 suggest that the prediction distribution generated by ordinary bagging has greater volatility than that of BBagging, as the range of most of the ordinary bagging results is larger than those of BBagging. The Wilcoxon test of the coefficient of variation of the forecast errors (see Table 8) reveals that the variation in the BBagging results is significantly smaller than the variation in the ordinary bagging results ($W = 781$, $p = 0.068$), which is consistent with the observation in the violin plots. The Wilcoxon test confirms that the difference is significant, especially in the variations in the short-run (one-step to three-steps ahead) forecasts ($W = 258$, $p = 0.001$). Consistent with the annual forecasts, the BBagging results are more concentrated than those of the ordinary bagging method in the quarterly forecasts, which means that the performance of BBagging has stronger validity. Overall, although the two bagging methods cannot improve the performance of GETS in quarterly forecasting, their performance for quarterly forecasting is consistent with their performance for annual forecasts. Specifically, although BBagging is not superior to ordinary bagging in terms of forecasting accuracy, its validity is much stronger.



Note. The red is the distribution of ordinary bagging and the blue is the distribution of BBagging forecasts.

Figure 4. Distributions of Quarterly Forecasts Using Ordinary Bagging and BBagging

Table 8 Coefficients of Variation – Quarterly Forecast Errors

Horizon	Australia		Canada		France	
	OB	BB	OB	BB	OB	BB
1	0.038	0.021	0.039	0.015	0.040	0.026
2	0.049	0.034	0.040	0.025	0.041	0.029
3	0.054	0.039	0.035	0.025	0.030	0.027
6	0.076	0.063	0.048	0.039	0.044	0.040
9	0.093	0.072	0.068	0.057	0.044	0.039
12	0.110	0.081	0.059	0.053	0.043	0.041
Horizon	Germany		UK		US	
	OB	BB	OB	BB	OB	BB
1	0.032	0.027	0.033	0.023	0.028	0.016
2	0.035	0.029	0.030	0.040	0.028	0.015
3	0.032	0.031	0.041	0.046	0.024	0.016

6	0.040	0.042	0.026	0.043	0.040	0.026
9	0.043	0.051	0.035	0.060	0.040	0.028
12	0.036	0.052	0.030	0.058	0.039	0.029

Abbreviation: BB, BBagging; OB, ordinary bagging.

5. Conclusion

This study presents the first attempt to introduce BBagging to tourism demand forecasting and investigates whether BBagging can improve the GETS model's annual and quarterly forecasts. Tourism demand for Hong Kong from six long-haul source markets (Australia, Canada, France, Germany, the UK, and the US) is used to compare the forecasts. Three forecasting methods are used to generate annual and quarterly forecasts: GETS, GETS combined with ordinary bagging, and GETS combined with BBagging. Their forecasting accuracy and variation are compared across the three methods for both annual and quarterly forecasts, which represent small and large samples, respectively.

This study reaches the following conclusions. First, both BBagging and ordinary bagging can improve the performance of the GETS model for small samples. However, their superiority is not evident in large samples. As the sample size increases, the superiority of the bagging methods, particularly BBagging, over the GETS procedure is not fully reflected. Second, although ordinary bagging and BBagging produce similar forecasting accuracy for small and large samples, BBagging has stronger validity than ordinary bagging. Given the same forecasting accuracy, the method that can generate concentrated and consistent forecasts will be preferred by decision makers. The BBagging method offers a robust option for forecasters seeking to predict tourism demand using fluctuating historical data. Decision makers can also use BBagging forecasts to inform their future planning and actions, as BBagging forecasts are more reliable than those generated by conventional one-off forecasting methods. As most tourism investments, such as airports and hotels, have a long construction period or high sunk costs, more reliable forecasts can help investors reduce the risk of forecasting failures and save financial resources. Due to the significant negative impacts of the social unrest in Hong Kong in 2019 and the outbreak of the COVID-19 pandemic in 2020 on its tourism industry, the empirical analyses in this study only use data from up to the end of 2018, which is a limitation of this study. However, given BBagging's capacity to handle strong variations in forecasts, future research could combine BBagging with GETS to forecast the recovery in tourism demand after COVID-19. To further improve its forecasting performance, subsequent research could focus on adding more information to the prior distributions of BBagging to better reflect the influence of historical data. Different averaging methods could also be introduced into the BBagging algorithm. Researchers could consider assigning different weights to the predictions or clustering the predictions to achieve better performance.

Acknowledgment

The authors acknowledge the financial support of The Hong Kong Polytechnic University (Grand No.: 5-ZJLP).

References

Athanasopoulos, G., Song, H., & Sun, J. A. (2018). Bagging in tourism demand modeling and forecasting. *Journal of Travel Research*, *57*(1), 52–68.

Bergmeir, C., Hyndman, R. J., & Benítez, J. M. (2016). Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation. *International Journal of Forecasting*, *32*(2), 303–312.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, *30*(4), 927–961.

Campos, J., Ericsson, N. R., & Hendry, D. F. (2005). General-to-specific modeling: An overview and selected bibliography. *FRB International Finance Discussion Paper*, (838).

Clyde, M. A., & Lee, H. (2001). Bagging and the Bayesian bootstrap. In T. Richardson and T. Jaakola (Eds.), *Artificial intelligence and statistics: Proceedings of the 8th International Workshop*. January 4–7, 2001, Morgan Kaufmann, Key West, Florida. (pp. 169–174).

Crouch, G. I. (1992). Effect of income and price on international tourism. *Annals of Tourism Research*, *19*(4), 643–664.

Fushiki, T. (2010). Bayesian bootstrap prediction. *Journal of Statistical Planning and Inference*, *140*(1), 65–74.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Hendry, D. F. (1995). *Dynamic econometrics.* Oxford University Press on Demand.

Inoue, A., & Kilian, L. (2008). How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation. *Journal of the American Statistical Association*, *103*(482), 511–522.

Kurogi, S., & Harashima, K. (2009). Improving generalization performance of bagging ensemble via Bayesian approach. In *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation-(CIRA)* (pp. 557–561). IEEE.

Lee, H. K., & Clyde, M. A. (2004). Lossless online Bayesian bagging. *Journal of Machine Learning Research*, *5*(Feb), 143–151.

Lee, T. H., & Yang, Y. (2006). Bagging binary and quantile predictors for time series. *Journal of Econometrics*, *135*(1–2), 465–497.

Li, G., Song, H., & Witt, S. F. (2005). Recent developments in econometric modeling and forecasting. *Journal of Travel Research*, *44*(1), 82–99.

Li, G., Wu, D. C., Zhou, M., & Liu, A. (2019). The combination of interval forecasts in tourism. *Annals of Tourism Research*, 75, 363–378.

Lin, V. S., Goodwin, P., & Song, H. (2014). Accuracy and bias of experts' adjusted forecasts. *Annals of Tourism Research*, *48*, 156–174.

Lin, V. S., Liu, A., & Song, H. (2015). Modeling and forecasting Chinese outbound tourism: An econometric approach. *Journal of Travel & Tourism Marketing*, *32*(1–2), 34–49.

Liu, A., Lin, V. S., Li, G., & Song, H. (2020). Ex ante tourism forecasting assessment. *Journal of Travel Research*, 1-12, doi:0047287520974456.

Liu, A., & Pratt, S. (2017). Tourism's vulnerability and resilience to terrorism. *Tourism Management*, *60*, 404–417.

Narayan, P. K. (2004). Fiji's tourism demand: The ARDL approach to cointegration. *Tourism Economics*, *10*(2), 193–206.

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(1), 3–26.

Oza, N. C., & Russell, S. (2001). Online bagging and boosting. In T. Richardson and T. Jaakola (Eds.), *Artificial intelligence and statistics: Proceedings of the 8th International Workshop*. January 4–7, 2001, Morgan Kaufmann, Key West, Florida. (pp. 105–112).

Page, S., Song, H., & Wu, D. C. (2012). Assessing the impacts of the global economic crisis and swine flu on inbound tourism demand in the United Kingdom. *Journal of Travel Research*, *51*(2), 142–153.

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.

Pesaran, H. M., Shin, Y., & Smith, R. J. (2001). Bounds testing approaches to the analysis of level relationships. *Journal of Applied Econometrics, 16*(3), 289–326.

Petropoulos, F., Hyndman, R. J., & Bergmeir, C. (2018). Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research*, *268*(2), 545–554.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, *9*(1), 130–134.

Song, H., Dwyer, L., Li, G., & Cao, Z. (2012). Tourism economics research: A review and assessment. *Annals of Tourism Research*, *39*(3), 1653–1682.

Song, H., & Fei, B. (2007). Modelling and forecasting international tourist arrivals to mainland China. *中国旅游研究 Journal of China Tourism Research*, *3*(1), 20–40.

Song, H., Gao, B. Z., & Lin, V. S. (2013). Combining statistical and judgmental forecasts via a web-based tourism demand forecasting system. *International Journal of Forecasting*, *29*(2), 295–310.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting: A review of recent research. *Tourism Management*, *29*(2), 203–220.

Song, H., & Lin, S. (2010). Impacts of the financial and economic crisis on tourism in Asia. *Journal of Travel Research*, *49*(1), 16–30.