Combination Forecasts of International Tourism Demand

Shujie Shen
University of Leeds, UK

Gang Li
University of Surrey, UK

Haiyan Song
The Hong Kong Polytechnic University, Hong Kong

Abstract

This study investigates the performance of combination forecasts in the context of international tourism demand. Five econometric and two time-series models are employed to generate individual forecasts. Six combination methods are then employed, and their forecasting performance evaluated, using data on UK outbound tourism demand in seven destination countries. The results suggest that combination forecasts, in general, outperform the best individual forecasts. More sophisticated such combination methods, which take the historical performance of individual forecasts into account, perform better than the simple average technique. The performance of combination forecasts is associated with the performance consistency of the individual forecasts they include, and the inclusion of up to three individual forecasts is most likely to result in accurate combination forecasts.

Keywords: Forecast combination; Demand forecasting; Forecasting horizon; Encompassing tests.

1 The paper should be cited as follows: Shen S, Li G, Song H. (2011) 'Combination forecasts of International tourism demand'. Annals of Tourism Research, 38 (1), pp. 72-89. doi: 10.1016/j.annals.2010.05.003

2 Shujie Shen, PhD is a research fellow at the Institute for Transport Studies, University of Leeds, UK; Gang Li (School of Management, University of Surrey, Guildford, United Kingdom, GU2 7XH; Email <g.li@surrey.ac.uk>), PhD is a senior lecturer in economics in the Faculty of Management and Law at the University of Surrey, UK; and Haiyan Song, PhD is chair professor and associate director of the School of Hotel and Tourism Management, The Hong Kong Polytechnic University, Hong Kong.
INTRODUCTION

Accurate forecasts of tourism demand provide crucial information for tourism policymaking in both the government and business sectors. Although many quantitative methods have been applied to forecasts of tourism demand, no single forecasting method has been found to outperform all others in all situations (Li, Song, and Witt, 2005). Empirical studies in other fields have shown that combining the forecasts obtained from single models can improve forecasting accuracy. This study evaluates the forecasting performance of different combination methods in the context of tourism demand.

COMBINATION FORECASTS OF TOURISM DEMAND

The past four decades has produced a substantial body of research on combination forecasts in the economics and business fields. The rationale for combining forecasts is that greater accuracy can be achieved by synthesising the information contained in different individual forecasts into a composite forecast (Bates and Granger, 1969; Winkler, 1989). Bunn (1989) notes that such combinations improve forecast accuracy by taking advantage of the availability of multiple information and computing resources, and defines this approach as ‘data-intensive’ forecasting (p. 162). Another principal motivation for combining forecasts is to avoid the difficulty and risks inherent in model selection. As Zhang (2003) states, “the final selected model is not necessarily the best for future uses due to many potential influencing factors such as sampling variation, model uncertainty and structure change. By combining different methods, the problem of model selection can be eased with little extra effort” (p. 160).

Although empirical research indicates that combining the forecasts of different models can improve forecast accuracy (see, for example, Chan, Stock, & Watson, 1999; Diebold & Pauly, 1990; Marcellino, 2002; Stock & Watson, 1999), few studies have included modern econometric models, such as the error-correction and time-varying-parameter (TVP) models, in their comparisons. Exceptions include the studies carried out by Shen, Li, and Song (2008), Song, Witt, Wong, and Wu (2009), and Wong, Song, Witt, and Wu (2007). As econometric models are based on different assumptions about the form that the relationship between variables takes (such as endogeneity in the vector autoregressive [VAR] model and exogeneity in most others) or employ different estimation methods (for example, the TVP model uses the Kalman filter algorithm), their model properties differ from one another and from time-series models. Bates and Granger (1969) suggest that the combination of models that contain independent information is most likely to improve forecast accuracy.

It is clear that it is useful to consider a variety of individual forecasting methods in any combination exercise. This study provides comprehensive empirical evidence that allows a comparison of the accuracy of combination versus individual forecasts. Seven single forecasting techniques are employed to generate individual forecasts. They include five econometric models, namely, the reduced autoregressive distributed lag model (READLM), the Wickens-Breusch (1988) error correction model (WB-ECM), the Johansen (1988) maximum likelihood error correction model (JML-ECM), the VAR model, and the time-varying-parameter model, and two time-series models, namely, the
seasonal naïve (no-change) model and the seasonal autoregressive integrated moving average (SARIMA) model. Since Bates and Granger (1969) first proposed their optimal method, a number of new combination methods have been developed (see the surveys of the forecast combination literature published by Clemen [1989], Diebold and Lopez [1996], Granger [1989], and Newbold and Harvey [2002]).

Despite this large body of literature on combination forecasts, however, the evidence on which is the best remains inconclusive. The presence of substantial bias in constituent forecasts prompted MacDonald and Marsh (1994) to employ ordinary least squares regression to combine exchange rate forecasts. The superiority of the regression method is supported by the work of Guerard (1987) and Holmen (1987). Mills and Stephenson (1985), Clemen (1986), Holden and Peel (1986), and Lobo (1991), in contrast, provide empirical evidence to show the superiority of other approaches over ordinary least squares regression. Some studies suggest that methods that weight better-performing forecasts more heavily are likely to perform better than the simple average (SA) combination technique, although there is a significant amount of empirical evidence to show that simple combination forecasts with equal weights outperform more sophisticated combination forecasts (e.g., Makridakis & Winkler, 1983; Stock & Watson, 2004).

LeSage and Magura (1992) and Deutsch, Granger, and Teräsvirta (1994) suggest that time-varying combination methods can improve forecast accuracy. According to He and Xu (2005), when certain individual methods employ nonlinear models, or when individual forecasts are derived on the basis of the non-linear functions of the information set, non-linear combination methods are more suitable (see Donaldson and Kamstra [1996] and Shi and Liu [1993] for empirical applications). One purpose of this study is thus to evaluate and compare the forecasting accuracy of different combination methods in an attempt to identify the optimal method of combining forecasts. The combination methods examined herein are the SA combination method, the variance-covariance combination (VACO) method, the discounted mean square forecast error (MSFE) method, the Granger-Ramanathan (GR) regression method, the shrinkage method, and the TVP forecast combination method, all of which have been explored in the fields of economics and business.

Due to the small sample size of the datasets adopted in this study, such methods as the artificial neural network method are not included. Empirical evidence is gathered on UK outbound tourism demand for seven major destinations. Although demand forecasting is of the utmost importance in tourism planning and policymaking for both the public and business sectors, applications of combination forecast methods remain rare in the tourism context. Exceptions include Chu (1998), Fritz, Brandon, and Xander (1984), Oh and Morzuch (2005), Shen et al. (2008), Song et al. (2009), and Wong et al. (2007). This study provides additional and more robust empirical evidence on the usefulness of forecast combinations in the tourism context by employing six different combination techniques across seven individual forecasting models on seven datasets and over five horizons.
Three advanced combination methods are introduced to the tourism forecast combination literature for the first time, thus providing tourism forecasting practitioners with new alternatives. The number of individual forecasting methods, and hence the number of combination options, presented here are amongst the highest in the general forecast combination literature, thereby providing more robust evidence to answer such questions as “How many individual forecasts should be combined?” and “Which forecasting methods should be combined?” The inclusion of a number of datasets helps us to examine whether a particular combination method performs consistently well and whether the properties of a forecasting series affect the performance of combination forecasts. In addition, encompassing tests are carried out to examine whether they contribute to the enhanced performance of these forecasts.

**Forecast Combination Methods**

**Simple Average Combination Method.** The SA combination method calculates composite forecasts by taking the arithmetic average of individual forecasts. Clemen (1989) concludes that the virtues of this method include impartiality, robustness, and a good track record in economic and business forecasting. It is thus a common choice in forecast combination studies and serves as a useful benchmark here. The method can be expressed as $$f_{ct} = \frac{1}{n} \sum_{i=1}^{n} f_{it}$$, where $$f_{ct}$$ denotes the combined forecast, $$f_{it}$$ is the ith forecast in time period t, and n is the number of forecasts to be combined.

**Variance-Covariance Method.** Bates and Granger (1969) introduce the VACO method. In the two-model forecast combination case, the combined forecasts are given as $$f_{ct} = w f_{1t} + (1-w) f_{2t}$$, where $$f_{ct}$$ is the combination forecast based on the individual forecasts of $$f_{1t}$$ and $$f_{2t}$$, and w and (1-w) are the weights assigned to $$f_{1t}$$ and $$f_{2t}$$, respectively. The weight that minimises the combined forecast variance is $$w^* = \frac{(\sigma_{12}^2 - \sigma_{11}^2)}{\sigma_{11}^2 + \sigma_{22}^2 - 2\sigma_{12}}$$, where $$\sigma_{11}^2$$ and $$\sigma_{22}^2$$ are the unconditional individual forecast error variances, and $$\sigma_{12}$$ is the covariance. In practice, a possible estimator of the combination weight is $$w_i = \frac{\sum_{t=1}^{T} e_{it}^2}{\sum_{t=1}^{T} e_{1t}^2 + \sum_{t=1}^{T} e_{2t}^2}$$, where $$e_{it}$$ and $$e_{2t}$$ are individual forecast errors, and T is the sample size (as $$\sigma_{11}^2$$, $$\sigma_{22}^2$$, and $$\sigma_{12}$$ are unknown). According to Fritz et al. (1984), the foregoing formula can be easily extended to include more than two individual forecasts, and the weights can be calculated by $$w_i = \frac{[\sum_{t=1}^{T} e_{it}^2]^{-1}}{\sum_{j=1}^{m} [\sum_{t=1}^{T} e_{jt}^2]^{-1}}$$.

**Granger and Ramanthan Regression Method.** The regression method developed by Granger and Ramanthan (1984) proceeds by regressing actual values on competing individual forecasts and a constant term, and then employing least squares parameter estimates to produce a combination forecast: $$f_{ct} = \tilde{\beta}_0 + \tilde{\beta}_1 f_{1t} + \tilde{\beta}_2 f_{2t} + ... + \tilde{\beta}_n f_{nt}$$, where...
\( f_{ct} \) represents a combined forecast based on a linear combination of \( k \) individual forecasts, \( f_i \) \((i = 1, 2, \ldots, n)\), and \( \hat{\beta}_i \) \((i = 0, 1, \ldots, n)\) denotes the least squares estimator based on observations up to time \( t-1 \), that is, the actual values at period \( t-1 \). The series of \( y_{t-1} \) is regressed against the individual forecasts, \( f_{i,t-1} \) \((i = 1, 2, \ldots, n)\), and a constant term to determine \( \hat{\beta}_i \) \((i = 0, 1, \ldots, n)\).

**Discounted Mean Square Forecast Error Method.** The discounted MSFE method was first proposed by Bates and Granger (1969) for a two-individual-forecast case and subsequently generalised by Newbold and Granger (1974) for an \( n \)-individual-forecast combination. The method makes use of the full sample, but weights recent observations more heavily (Diebold & Pauly, 1987; Winkler & Makridakis, 1983). The combination of \( n \) individual forecasts for period \( t \) is given as \( f_{ct} = \sum_{i=1}^{n} w_i f_i \), where \( f_i \) is the forecast for period \( t \) from forecasting method \( i \), \( w_i \) is the weight assigned to individual forecast \( f_i \), and \( n \) is the number of individual forecasts. The weight of the discounted MSFE of the combined forecasts is specifically defined as \( w_i = \phi^{-1} \sum_{j=1}^{n} \phi^{-1} \), where 
\[
\phi = \sum_{s=1}^{t} \alpha^{t-s} (y_s - f_s^i)^2, \quad \alpha \text{ is a discounting factor with } 0 < \alpha \leq 1, \text{ and } y_s \text{ is the actual value for period } s.
\]

**Shrinkage Method.** Drawing on Clemen and Winkler (1986), Diebold and Pauly (1990) employ Bayesian shrinkage techniques to allow the incorporation of varying degrees of prior information into the estimation of combination weights. In this shrinkage method, the least square weights and arithmetic mean emerge as the two extreme cases for the posterior mean. The actual posterior mean combination weights are a matrix-weighted average of those for the two extreme cases. The exact location depends on prior precision, which can be estimated from the data using an empirical Bayesian procedure. Such procedures, which employ shrinkage towards a measure of central tendency (e.g., the arithmetic mean), are increasingly playing a role in forecast combinations.

Although the combination weights are coaxed towards the arithmetic mean, the data are still allowed to speak when (and if) they have something to say. The shrinkage method computes the weights as an average of the recursive ordinary least squares estimator of the weights based on the GR method and equal weighting, that is, \( w_i = \lambda \hat{\beta}_i + (1-\lambda)(1/n) \), where \( \hat{\beta}_i \) is the \( i \)th estimated coefficient from a recursive ordinary least squares regression of \( y_t \) against \( f_{1t}, \ldots, f_{nt} \) for \( t = 1 \) to \( T \), and \( \lambda = \max\{0,1-\kappa [n/(T-1-n)]\} \), where \( \kappa \) is a constant that controls the amount of shrinkage towards equal weighting and \( \kappa \) takes a value between 0 and 1. A larger \( \kappa \) corresponds to more shrinkage towards equal weighting.

**Time-Varying-Parameter Combination Method with the Kalman Filter.** This method utilises the Kalman filter algorithm to estimate the coefficients in the combined
regression, which are assumed to follow a random walk process. It has been used by Sessions and Chatterjee (1989), LeSage and Magura (1992), and Stock and Watson (2004). The TVP combination method begins with the GR regression model with time-varying parameters, 

$$f_{t} = \beta_{0t} + \beta_{1t}f_{t-1} + \beta_{2t}f_{t-2} + \ldots + \beta_{kt}f_{t-k} + \epsilon_t \quad \text{and} \quad \beta_{it} = \beta_{i,t-1} + \eta_{it},$$

where $\eta_{it}$ is independent and identically distributed and is uncorrelated with $\epsilon_t$. The Kalman filter approach also facilitates real-time parameter ‘updating’ and can readily handle both stationary (e.g., autoregressive moving average) and non-stationary (e.g., integrated autoregressive moving average) parameter drifts (Diebold & Pauly, 1987).

**Data and Forecasting Models**

This empirical study focuses on UK outbound tourism demand for seven major destinations, Spain, France, Greece, Italy, Australia, Canada and the USA, which are the leading short- and long-haul markets for such outbound tourism. The tourism demand function for this study takes the following general form:

$$LQ_{it} = f(LY_{it}, LP_{it}, LQ_{it-1}, \text{Dummies}),$$

where $L$ indicates a logarithm; $Q_{it}$ is UK outbound tourism demand measured by quarterly tourist arrivals in destination country $i$; $Y_{it}$ is tourist income measured by real gross domestic product in the UK in constant prices (1995 = 100); and $P_{it}$ represents the relative tourism price for each individual destination country $i$, which is calculated by dividing the price (measured by the consumer price index, $CPI_{it}$) in each destination by that in the UK ($CPI_{UK}$) and then adjusting by the relevant exchange rates ($EX_{it}$ and $EX_{UK}$), as follows.

$$P_{it} = \frac{CPI_{it}/EX_{it}}{CPI_{it}/EX_{it}},$$

(1)

where $P_{ist}$ represents the relative substitute price in competing destinations, measured by the weighted average price index amongst the main alternative destinations for each destination country relative to the tourism prices in the origin country, the UK, with the shares of tourist arrivals from the UK in these alternatives serving as the weights.

The short-haul (France, Spain, Greece and Italy) and long-haul destinations (Australia, Canada and the USA) are regarded as substitutes for one another within each group. New Zealand is also added to the long-haul substitution set, as it is one of the most popular such destinations for UK tourists. Hence, the substitute destinations for the USA are Canada, Australia and New Zealand; those for Canada are the USA, Australia and New Zealand; and those for Australia are the USA, Canada and New Zealand. Divisekera (2003) makes the case for choosing these countries as substitutions. In the case of France, the substitute price is defined as

$$P_{sp} = \frac{P_{sp} \cdot TOU_{sp} + P_{st} \cdot TOU_{st} + P_{gr} \cdot TOU_{gr}}{TOU_{sp} + TOU_{st} + TOU_{gr}}.$$  

(2)

Three dummy variables are included to capture the effects of one-off events on UK outbound tourism demand: DUM86 represents the severe decline in world oil prices in 1986 (DUM86 = 1 in 1986Q2 and 1986Q3, and 0 otherwise); DUM90 captures the effects of Iraq’s invasion of Kuwait in 1990 (DUM90 = 1 in 1990Q3 and 1990Q4, and 0 otherwise); DUM86 represents the severe decline in world oil prices in 1986 (DUM86 = 1 in 1986Q2 and 1986Q3, and 0 otherwise); DUM90 captures the effects of Iraq’s invasion of Kuwait in 1990 (DUM90 = 1 in 1990Q3 and 1990Q4, and 0 otherwise);
otherwise); and DUM91 those of the Gulf War in 1991 (DUM91 = 1 in 1991Q1, 1991Q2 and 1991Q3, and 0 otherwise). Seasonal dummies are also included in some of the models to capture the effects of seasonality: \( D_{i,t} \) \((i = 2,3,\ldots,s) = 1\) if time \( t \) corresponds to season \( s \), and \( D_{i,t} = 0 \) otherwise \((s = 4\) here, as the study adopts quarterly time series).

The data cover the 1984Q1-2004Q4 period. Those on real gross domestic product, exchange rates and consumer price indexes are obtained from the International Monetary Fund’s *International Financial Statistics Yearbook*, and those on tourist arrivals from the United Nations World Tourism Organization’s *Tourism Statistical Yearbook*. Five modern econometric models commonly used in the tourism demand forecasting literature are employed to generate individual *ex post* forecasts: the READLM, WB-ECM, JML-ECM, VAR model and TVP model. The model specifications follow Song, Witt and Jensen (2003). In addition, two widely used univariate time-series models, the seasonal naïve and SARIMA models, are included as benchmarks for comparison purposes. Their specifications are in line with Witt and Witt (1992) and Pankratz (1983). As quarterly data are employed, the time series is likely to exhibit seasonality. The HEGY test developed by Hylleberg, Engle, Granger and Yoo (1990) is thus adopted to test for seasonal and non-seasonal unit roots. Two types of seasonality are relevant here: stochastic and deterministic seasonality. The former requires seasonal differencing, and the latter seasonal dummies, in the model estimation.

**Empirical Results**

**Performance of Individual Forecasts.** The individual forecasts obtained from the seven forecasting models are generated at one- to four- and eight-quarter-ahead horizons. To ensure consistency with previous tourism forecasting studies, our accuracy comparison is based on frequently used error measures: the mean absolute percentage error (MAPE) and root mean squared percentage error. As these two measures result in highly consistent individual and combination forecast performance evaluations, only the MAPE results are reported here. The performance rankings of the alternative models based on the MAPE are presented in Table 1. Due to space constraints, the forecasting performance of individual models across the destination countries is not reported here, but this information is available upon request.

Overall, the JML-ECM is found to be superior to all of the other models in forecasting UK outbound tourist arrivals in the seven major destination countries. The time-series models are ranked second (the naïve model) and third (the SARIMA model). Although the naïve model closely follows the JML-ECM in terms of the overall forecast errors across countries, the two exhibit different levels of performance in individual cases. For example, the latter outperforms the former in the cases of Spain and the USA, whereas the opposite is the case with Canada and France. The WB-ECM performs best in the case of Italy, and is ranked fourth overall. The READLM generates the least accurate forecasts. Although previous empirical studies (e.g., Song & Witt, 2000; Song & Wong, 2003) suggest that the TVP model performs well in annual tourism demand forecasting, in this study, it was outperformed by most of the other models when quarterly data were involved.
Table 1 Forecast Accuracy of Individual Models at Different Forecasting Horizons

<table>
<thead>
<tr>
<th>Horizon</th>
<th>NAÏVE</th>
<th>SARIMA</th>
<th>READLM</th>
<th>WB-ECM</th>
<th>JML-ECM</th>
<th>TVP</th>
<th>VAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 quarter</td>
<td>2.632 (2)</td>
<td>2.890 (3)</td>
<td>4.195 (6)</td>
<td>3.370 (4)</td>
<td>2.572 (1)</td>
<td>4.484 (7)</td>
<td>3.983 (5)</td>
</tr>
<tr>
<td>2 quarters</td>
<td>2.608 (2)</td>
<td>2.857 (3)</td>
<td>3.840 (5)</td>
<td>3.642 (4)</td>
<td>2.588 (1)</td>
<td>4.941 (7)</td>
<td>4.142 (6)</td>
</tr>
<tr>
<td>3 quarters</td>
<td>2.629 (1)</td>
<td>2.898 (3)</td>
<td>4.136 (5)</td>
<td>3.791 (4)</td>
<td>2.804 (2)</td>
<td>4.554 (7)</td>
<td>4.137 (6)</td>
</tr>
<tr>
<td>4 quarters</td>
<td>2.623 (1)</td>
<td>2.837 (2)</td>
<td>4.511 (7)</td>
<td>3.953 (5)</td>
<td>2.862 (3)</td>
<td>3.217 (4)</td>
<td>4.136 (6)</td>
</tr>
<tr>
<td>8 quarters</td>
<td>3.515 (3)</td>
<td>2.680 (1)</td>
<td>6.172 (7)</td>
<td>4.531 (4)</td>
<td>3.101 (2)</td>
<td>4.630 (5)</td>
<td>5.362 (6)</td>
</tr>
<tr>
<td>Overall</td>
<td>2.801 (2)</td>
<td>2.833 (3)</td>
<td>4.571 (7)</td>
<td>3.857 (4)</td>
<td>2.786 (1)</td>
<td>4.365 (6)</td>
<td>4.352 (5)</td>
</tr>
</tbody>
</table>

Note: SARIMA denotes the seasonal autoregressive integrated moving average model, READLM the reduced autoregressive distributed lag model, WB-ECM the Wickens-Breusch (1988) error correction model, JML-ECM the Johansen (1988) maximum likelihood error correction model, TVP the time-varying-parameter model, and VAR the vector autoregressive model. Figures in brackets indicate the rank of the forecasting method; accuracy evaluation is based on the MAPE forecast error measure.

The treatment of seasonality in the data series may provide an explanation for the relatively poor performance achieved by the READLM, TVP model and VAR model. As noted, seasonality in a time series is regarded as either deterministic or stochastic in nature. When it is stochastic, the time series needs to be seasonally differenced to account for seasonal unit roots, whereas seasonal dummies should be used to capture deterministic seasonality. The specifications of the JML-ECM and WB-ECM are based on the results of the seasonal unit root tests. The seasonal naïve and SARIMA models assume seasonal unit roots at seasonal frequencies, an assumption that is consistent with the results of the seasonal unit root tests, thus indicating that UK outbound tourist arrivals exhibit stochastic seasonality. The other three models, in contrast, use seasonal dummies to account for the seasonality in tourism demand, although such simplification is incapable of reflecting the dynamic nature of this seasonality. Moosa and Kennedy (1998) draw a similar conclusion.

Our empirical results suggest that no single forecasting method is able to outperform all others in all situations. For example, as noted, the JML-ECM outperforms its competitors in terms of overall performance, but is outperformed by the naïve model in two of the seven country cases. It is thus believed that combining the forecasts generated by these individual methods may be a favourable option. An important observation is that the level of performance achieved by the individual forecasting models varies across the seven destinations. Taking the one-quarter-ahead forecasts as an example, Figure 1 shows that the MAPE values amongst the seven individual models are similar in the cases of France, Spain, the USA and Australia, whereas considerable variations are observed in the cases of Canada, Greece and Italy. In the case of Italy, for example, the MAPE of the least accurate model (the TVP model) is two-and-a-half times that of the most accurate (the WB-ECM). Similar patterns also appear over the longer forecasting horizons.

Further examination of the seven demand series suggests that the demand for tourism to France, Spain, Australia and the USA grew more steadily over the sample period, thus making it relatively easy to forecast future trends, regardless of the method employed.
Demand for tourism to Canada, Greece and Italy, in contrast, was characterised by greater fluctuation, and is therefore more difficult to predict, at least with some of the methods. Accordingly, we separate the destinations into two groups: France, Spain, Australia and the USA are designated Group A, and Canada, Greece and Italy Group B. The consistency/inconsistency in the performance among individual models in the case of a particular destination, or the difficulty in predicting future demand for that destination, may affect the results of the combination methods of which they are part, as is further discussed in the following section.

![Figure 1. One-step-ahead Forecast MAPEs of Individual Models across Destination Cases](image)

**Performance of Combination Forecasts and Comparison with Individual Forecasts.**

The combination forecasts are created from the individual forecasts generated by the seven single forecasting approaches. The individual out-of-sample forecasts at each horizon are divided into two sub-samples, with the first 20 observations employed for the construction of the combination weights and the remainder for a post-sample evaluation of the combined forecasts. The recursive scheme adopted to calculate the weights is as follows. The first 20 observations are used to obtain the weights, which are subsequently used to calculate the combination forecast for the following period. The data are then updated with one additional observation to obtain the new weights for the calculation of the next combination forecast. This procedure is repeated until the final combination forecast is obtained. The discounted MSFE combination forecasts are computed using three values of discounting factor $\alpha$: 0.95, 0.9 and 0.85. The shrinkage forecasts are evaluated for $\kappa = 0.25, 0.5, 0.75$ and 1.

For each destination country, seven individual forecast series are available for combination at each forecasting horizon, and all possible combinations of the seven methods are considered. Thus, for each combination method applied, 120 combinations
are carried out at each forecasting horizon for all of the destination countries. The percentage of superior combination forecasts, i.e., the proportion of combination forecasts that are more accurate than the best component individual forecasts (based on the MAPE and root mean squared percentage error, respectively) for all 120 combinations, is employed to evaluate forecasting performance. The performance of the combination and individual forecasts is first compared and evaluated across the different destination countries (Table 2) and then across the different forecasting horizons (Table 3). For the discounted MSFE and shrinkage combination forecasts, the results of the best options (i.e., $\alpha = 0.85$ and $\kappa = 1$, respectively) are reported.

### Table 2 Percentages of Superior Combination Forecasts in Different Cases

<table>
<thead>
<tr>
<th>Destination</th>
<th>SA</th>
<th>VACO</th>
<th>GR</th>
<th>TVP</th>
<th>Shrink1</th>
<th>Dis.85</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>99.50 (3)</td>
<td>100.00 (1)</td>
<td>99.50 (3)</td>
<td>83.67 (4)</td>
<td>99.83 (2)</td>
<td>100.00 (1)</td>
</tr>
<tr>
<td>Spain</td>
<td>91.00 (3)</td>
<td>93.17 (2)</td>
<td>77.83 (5)</td>
<td>50.00 (6)</td>
<td>85.83 (4)</td>
<td>93.83 (1)</td>
</tr>
<tr>
<td>Australia</td>
<td>76.33 (5)</td>
<td>83.33 (4)</td>
<td>95.50 (1)</td>
<td>72.67 (6)</td>
<td>87.83 (2)</td>
<td>87.17 (3)</td>
</tr>
<tr>
<td>USA</td>
<td>62.83 (5)</td>
<td>73.33 (2)</td>
<td>85.67 (1)</td>
<td>67.67 (4)</td>
<td>57.83 (6)</td>
<td>71.00 (3)</td>
</tr>
<tr>
<td>Canada</td>
<td>17.50 (5)</td>
<td>36.83 (3)</td>
<td>50.67 (1)</td>
<td>10.50 (6)</td>
<td>32.00 (4)</td>
<td>37.17 (2)</td>
</tr>
<tr>
<td>Greece</td>
<td>53.83 (1)</td>
<td>29.00 (3)</td>
<td>8.33 (5)</td>
<td>4.50 (6)</td>
<td>12.67 (4)</td>
<td>36.33 (2)</td>
</tr>
<tr>
<td>Italy</td>
<td>12.17 (5)</td>
<td>29.83 (3)</td>
<td>28.83 (4)</td>
<td>6.50 (6)</td>
<td>44.67 (1)</td>
<td>32.00 (2)</td>
</tr>
</tbody>
</table>

Note: SA denotes the simple average combination method, VACO the variance-covariance method, GR the Granger-Ramanathan regression method, TVP the time-varying-parameter combination method, Shrink1 the shrinkage method with $\kappa = 1$, and Dis.85 the discounted MSFE method with $\alpha = 0.85$. Figures in brackets indicate the rank of the combination method in each destination case; accuracy evaluation is based on the MAPE forecast error measure.

### Table 3 Percentages of Superior Combination Forecasts at Different Forecasting Horizons

<table>
<thead>
<tr>
<th>Horizon</th>
<th>Summary method</th>
<th>SA</th>
<th>VACO</th>
<th>GR</th>
<th>TVP</th>
<th>Shrink1</th>
<th>Dis.85</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 quarter</td>
<td>Average</td>
<td>62.62 (2)</td>
<td>60.60 (4)</td>
<td>66.90 (1)</td>
<td>43.93 (6)</td>
<td>59.88 (5)</td>
<td>62.26 (3)</td>
</tr>
<tr>
<td></td>
<td>Aggregate</td>
<td>79.17 (5)</td>
<td>93.33 (3)</td>
<td>94.17 (2)</td>
<td>10.83 (6)</td>
<td>86.67 (4)</td>
<td>97.50 (1)</td>
</tr>
<tr>
<td>2 quarters</td>
<td>Average</td>
<td>48.57 (4)</td>
<td>50.83 (3)</td>
<td>56.79 (1)</td>
<td>28.69 (6)</td>
<td>43.33 (5)</td>
<td>51.67 (2)</td>
</tr>
<tr>
<td></td>
<td>Aggregate</td>
<td>56.67 (5)</td>
<td>88.33 (2)</td>
<td>80.83 (3)</td>
<td>9.17 (6)</td>
<td>78.33 (4)</td>
<td>92.50 (1)</td>
</tr>
<tr>
<td>3 quarters</td>
<td>Average</td>
<td>62.38 (4)</td>
<td>68.81 (2)</td>
<td>66.55 (3)</td>
<td>40.83 (6)</td>
<td>61.79 (5)</td>
<td>71.90 (1)</td>
</tr>
<tr>
<td></td>
<td>Aggregate</td>
<td>75.83 (5)</td>
<td>95.83 (2)</td>
<td>81.67 (4)</td>
<td>9.17 (6)</td>
<td>91.67 (3)</td>
<td>98.33 (1)</td>
</tr>
<tr>
<td>4 quarters</td>
<td>Average</td>
<td>63.93 (5)</td>
<td>69.05 (3)</td>
<td>64.64 (4)</td>
<td>46.19 (6)</td>
<td>70.83 (1)</td>
<td>70.00 (2)</td>
</tr>
<tr>
<td></td>
<td>Aggregate</td>
<td>73.33 (5)</td>
<td>80.00 (3)</td>
<td>90.83 (1)</td>
<td>20.83 (6)</td>
<td>77.50 (4)</td>
<td>90.00 (2)</td>
</tr>
<tr>
<td>8 quarters</td>
<td>Average</td>
<td>57.62 (5)</td>
<td>68.93 (2)</td>
<td>63.93 (4)</td>
<td>51.19 (6)</td>
<td>64.64 (3)</td>
<td>70.95 (1)</td>
</tr>
<tr>
<td></td>
<td>Aggregate</td>
<td>60.00 (5)</td>
<td>75.83 (2)</td>
<td>69.17 (4)</td>
<td>26.67 (6)</td>
<td>71.67 (3)</td>
<td>84.17 (1)</td>
</tr>
</tbody>
</table>

Note: SA denotes the simple average combination method, VACO the variance-covariance method, GR the Granger-Ramanathan regression method, TVP the time-varying-parameter combination method, Shrink1 the shrinkage method with $\kappa = 1$, and Dis.85 the discounted MSFE method with $\alpha = 0.85$. Figures in brackets indicate the rank of the combination method at each forecasting horizon; accuracy evaluation is based on the MAPE forecast error measure.
Comparison across Destination Countries. Table 2 summarises the average percentage of the superior combination forecasts over the five horizons under study in each destination case and by each combination method. Two distinctive patterns can be observed with regard to the performance of the combination forecasts. For the four destination countries in Group A (i.e., France, Spain, Australia and the USA), these forecasts perform overwhelmingly well, with each generating at least 50% superior forecasts. Hence, for at least half of the combination options, the combined forecasts outperform the best component forecasts, regardless of which combination method is used and which Group A destination is concerned. With regard to the best technique for each destination, in at least 85% of the cases the combination forecasts outperform the best component forecasts. The most successful case is France. Although the TVP combination method does not perform as well as the others in this case, nearly 84% of the combination results, on average, remain superior to the best component forecasts across all horizons.

However, the combination methods do not perform as well for the Group B countries, although, in the cases of Greece and Canada, the forecasts generated by the best of these methods outperform the best component forecasts more than half of the time. The relatively poor performance of the combination forecasts in Group B is due to the divergence in performance of the individual models. In the case of Canada, for which one or two of the individual forecasting models generate markedly inferior forecasts, a combination of methods is unlikely to improve forecasting accuracy. Alternatively, when one or two generate markedly superior forecasts (such as in the case of Italy), it is difficult for combination forecasts to outperform them. In the Group A cases, in contrast, the performance of the different individual forecasting models is highly consistent. Thus, when the models are combined, diversification gains are achieved.

These findings are consistent with the general forecast combination literature (e.g., Teräsvirta, Van Dijk, & Medeiros, 2005). The accuracy rankings of the six combination methods in each destination case suggest that no single such method outperforms the others in all cases. The discounted MSFE method with a discounting factor of 0.85 and the VACO method perform most consistently well and are both ranked above average in all but one case. The GR regression method exhibits the best performance in the cases of Australia, the USA and Canada, but its performance is ranked the second poorest in the cases of Greece and Spain. The SA method and the shrinkage method with $\kappa = 1$ (Shink1) outperform the competing models in the cases of Greece and Italy, respectively. Finally, the TVP combination method performs worst, being ranked at the bottom in five of the seven cases.

Comparison over Forecasting Horizons. The combination forecasts are evaluated at each forecasting horizon in two ways on the basis of their performance in each destination case: (a) calculated simply as the average percentage of superior combination forecasts across the seven cases; and (b) as the average MAPE amongst the seven cases for both the individual and combination forecasts, with comparison then made amongst them and the percentage of superior combination forecasts determined. The difference between these methods of calculation is the weighting scheme (i.e., equal versus
differential weights) assigned in the individual destination cases. Both are reported in Table 3 and denoted as ‘average’ and ‘aggregate’, respectively. It can be seen from this table that the combination forecasts perform consistently over the forecasting horizons except for the eight-quarter-ahead horizon. The aggregate method suggests that the performance of these forecasts over the shorter horizons (i.e., the one- to four-quarter-ahead horizons) is superior to that over the longer horizon (i.e., the eight-quarter-ahead horizon), except for the TVP combination method. The average method exhibits no clear pattern.

With regard to the relative performance of the six combination techniques, as evaluated by the aggregate method, the discounted MSFE method \((\alpha = 0.85)\) appears to be the best and most consistent over the five horizons, followed by the VACO and GR methods, outperforming all competing models over all horizons except for the four-quarter-ahead horizon. The SA method performs well at the one-quarter-ahead horizon, as judged by the average method, but this good performance is not maintained as the horizon is extended. The TVP combination method is outperformed by the others over all horizons. Its unsatisfactory performance is associated with the very short sub-samples used to calculate the optimal weights. Overall, with the exception of the TVP combination method, all of the five combination methods perform well over the forecasting horizons under study. In 48 of the 50 cases shown in Table 3, the percentages of the superior combination forecasts based on the five combination methods are greater than 50%. These results suggest that the length of the forecasting horizon has no significant effect on the performance of combination forecasts, at least as far as short-term forecasts (up to a year) are concerned. Thus this study provides empirical evidence to show that a combination of forecasts improves the accuracy of tourism demand forecasting.

In sum, the comparisons across the different destination cases and over the different forecasting horizons allow us to conclude that, of the six combination methods considered herein, the discounted MSFE method \((\alpha = 0.85)\) performs best and most consistently, followed by the VACO and GR regression methods. These results suggest that more sophisticated combination methods, that is, those that take the historical performance of individual forecasts into account, perform better than the SA method, a finding that is consistent with some of the general forecast combination literature, including the studies carried out by Holden and Peel (1986) and Bischoff (1989). With regard to the shrinkage method, the greater the shrinkage is, the more accurate the combination results are. With regard to the discounted MSFE, the lower the value of the discounting factor (i.e., the more discounting there is), the better the performance is.

**Encompassing Tests in Forecast Combination Practice.** Encompassing tests are employed to evaluate whether competing forecasts can be fruitfully combined to produce a forecast that is superior to individual forecasts (Fang, 2003). Two such tests, the regression-based F-test and the modified Diebold-Mariano-type test, are carried out in this study to identify the individual forecasts that can be most fruitfully combined (for details and justifications of these tests, see Harvey and Newbold [2000]). The F-test is applied to the one-quarter-ahead forecasts and the modified Diebold-Mariano-type test to
Individual forecasts are combined when the F-test results show none of them to encompass its competitors. The effect of the encompassing tests on the combination forecasts is measured by the percentage improvement in the superior such forecasts. With the exception of the TVP combination method (4.85% improvement), these tests result in a greater than 5% improvement in the performance of the combination forecasts. Due to space constraints, the percentage improvement in each combination method is unreported here, but these results are available upon request. It can be concluded that encompassing tests help to improve the performance of combination forecasts and that it is thus necessary to conduct such tests before individual forecasts are combined.

**Which and How Many Forecasting Models Should Be Combined?** In this section, we further discuss the relationship between the performance of individual forecasts and that of combination forecasts. Taking the best performing combination method (i.e., the discounted MSFE with \( \alpha = 0.85 \)) as an example, Table 4 presents the percentages of the individual forecasts in the superior combinations across the seven destination cases. These figures are the overall percentages over the five forecasting horizons. Consistent with the results presented in Table 2, two distinct patterns can be observed between Groups A and B. For the Group A countries, all seven individual forecasting models appear frequently in the superior combinations, although slight differences can be seen between the France-Spain and Australia-USA pairs. In the former pair, the frequencies are either identical or nearly so amongst all of the individual models regardless of their performance, whereas for the latter, the best performing model (the JML-ECM in both cases) appears most frequently, and the least accurate methods (the WB-ECM model, respectively) the least frequently, in the superior combinations.

An opposite trend is observed for the three destinations in Group B. The best performing models always make the least frequent appearances in the superior combinations, which indicates that once an individual forecast has performed inconsistently, the combination of a strongly superior and strongly inferior forecast is unlikely to outperform a superior component forecast. The contrast between Groups A and B suggests that combining the best individual forecasts does not always result in a superior combination forecast. Table 4 provides further evidence to confirm that consistency in the performance of individual forecasts affects the performance of their combination.

Many forecast combination studies include only two component forecasts. Although Bates and Granger (1969) suggest that “there is every reason to combine more than two forecasts (where they exist)” (p. 461). It is thus natural to ask whether a greater number of components leads to a better combination forecast and, if so, whether there is an
optimal such number. Table 5 presents empirical evidence gleaned from our investigation of whether the number of component forecasts affects the performance of combination forecasts. The best performing forecast amongst the seven individual forecasts and 120 combination options is identified in each of 60 comparisons (i.e., five forecasting horizons, six combination methods and two error measures; 5 x 6 x 2 = 60) for each destination case. The distribution of the 60 best performing forecasts amongst the one- (i.e., no combination) to seven-component combinations is then summarised.

Table 4 Frequencies of Individual Forecasts in Superior Combination Forecasts: Discounted MSFE ($\alpha = 0.85$)

<table>
<thead>
<tr>
<th>Country</th>
<th>NAIVE</th>
<th>SARIMA</th>
<th>READL</th>
<th>WB-ECM</th>
<th>JML-ECM</th>
<th>TVP</th>
<th>VAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>52.5</td>
<td>1</td>
<td>52.5</td>
<td>7</td>
<td>52.5</td>
<td>4</td>
<td>52.5</td>
</tr>
<tr>
<td>Spain</td>
<td>53.7</td>
<td>5</td>
<td>53.1</td>
<td>1</td>
<td>52.6</td>
<td>7</td>
<td>53.3</td>
</tr>
<tr>
<td>Australia</td>
<td>55.0</td>
<td>2</td>
<td>53.8</td>
<td>4</td>
<td>52.6</td>
<td>5</td>
<td>52.2</td>
</tr>
<tr>
<td>USA</td>
<td>57.4</td>
<td>4</td>
<td>52.5</td>
<td>2</td>
<td>49.1</td>
<td>5</td>
<td>52.3</td>
</tr>
<tr>
<td>Canada</td>
<td>23.0</td>
<td>1</td>
<td>54.2</td>
<td>2</td>
<td>63.8</td>
<td>5</td>
<td>38.3</td>
</tr>
<tr>
<td>Greece</td>
<td>42.2</td>
<td>2</td>
<td>45.2</td>
<td>3</td>
<td>69.8</td>
<td>6</td>
<td>57.3</td>
</tr>
<tr>
<td>Italy</td>
<td>70.4</td>
<td>4</td>
<td>52.6</td>
<td>3</td>
<td>65.8</td>
<td>6</td>
<td>18.0</td>
</tr>
</tbody>
</table>

Note: SARIMA denotes the seasonal autoregressive integrated moving average model, READLM the reduced autoregressive distributed lag model, WB-ECM the Wickens-Breusch (1988) error correction model, JML-ECM the Johansen (1988) maximum likelihood error correction model, TVP the time-varying-parameter model, and VAR the vector autoregressive model. Accuracy evaluation is based on the MAPE forecast error measure.

Table 5 Frequencies of Best Performing Combinations According to the Number of Components

<table>
<thead>
<tr>
<th>No. of component models</th>
<th>France</th>
<th>Spain</th>
<th>Australia</th>
<th>USA</th>
<th>Canada</th>
<th>Greece</th>
<th>Italy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1.7</td>
<td>8</td>
<td>13.3</td>
<td>26</td>
<td>43.3</td>
<td>27</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>17</td>
<td>30</td>
<td>50.0</td>
<td>33</td>
<td>55.0</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>40.0</td>
<td>14</td>
<td>23.3</td>
<td>13</td>
<td>21.7</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>23.3</td>
<td>9</td>
<td>15.0</td>
<td>13</td>
<td>21.7</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>5.0</td>
<td>7</td>
<td>11.7</td>
<td>5</td>
<td>8.3</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>5.0</td>
<td>5</td>
<td>8.3</td>
<td>3</td>
<td>5.0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1.7</td>
<td>1</td>
<td>1.7</td>
<td>1</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td>100</td>
<td>60</td>
<td>100</td>
<td>60</td>
<td>100</td>
<td>60</td>
</tr>
</tbody>
</table>

The highest frequencies of the best combination forecasts appear when there are three (in the Australia, Canada, France and Spain cases) or two components (in the Greece, Italy and USA cases), which suggests that combinations of up to three individual forecasts are most likely to result in accurate combination forecasts. Furthermore, in the cases of Australia, France and Spain, the best combination forecasts always outperform
the best individual forecasts. Therefore, for these cases, a combination of three individual forecasts is likely to generate not only the best combination forecast, but also the best overall forecast. The results presented in Table 5 also suggest that a combination of more than five individual forecasts is unlikely to be effective because of the greater likelihood that the information contained within them will overlap. Hence, including all forecasts in a single combination (Bates and Granger, 1969) is not advised.

CONCLUSION

This study evaluates and compares the empirical performance of various combination methods in the context of tourism demand forecasting. The individual forecasts are generated from five modern econometric models and two time-series models. In addition to the SA combination method, more sophisticated combination methods, including the VACO combination method, the discounted MSFE method, the GR regression method, the shrinkage method and the TVP forecast combination method, are employed. The effects of encompassing tests on the performance of the combination forecasts are also examined. The empirical results suggest that, overall, combination forecasts can improve forecast accuracy, as they are superior to the best individual forecasts. It is also found that the length of the forecasting horizon appears to have no significant effect on the performance of combination forecasts and that consistency in the performance of individual forecasting models is positively related to the performance of the forecasts that result from their combination.

Moreover, the more sophisticated combination forecasts, such as the discounted MSFE and VACO methods, which take the historical performance of individual forecasts into consideration, perform better than SA combination forecasts. The discounted MSFE method is found to be the best overall combination method in this study and the TVP combination method the worst. The performance of the latter is most likely due to the relatively small sample size employed, rather than to weaknesses with the method itself. We intend to pursue further examination of its performance with a sufficiently long time series in future research. In addition, it is found that encompassing tests improve the accuracy of combination forecasts, and thus it is necessary to conduct these tests before combining individual forecasts. Finally, it is found that combinations of up to three individual forecasts are likely to generate the best results.

The findings of this study have significant implications for both government agencies and the tourism industries. The former rely on accurate forecasts of tourism demand to guide their decisions about tourism-related infrastructure developments such as airports, highways and rail networks. In addition, appropriate government tourism policies and effective promotional strategies by destination management organisations require accurate predictions of future tourism growth, both at the aggregate level and based on individual source markets. Accurate forecasts of tourism demand are also crucial to tourism-related businesses, such as those in the transport, hospitality and visitor attraction sectors, which rely on them for successful capacity management (such as staffing and inventory control) and sound decisions about investment in facilities and equipment and business expansion (such as new hotels and expanded theme parks). When these policymakers and business managers are faced with more than one set of forecasts generated by different individual models, their best option is to combine the individual forecasts instead of choosing those produced by one model.
Different sets of forecasts based on different models may represent different prospects for further tourism growth. Without a strong belief in the likelihood of a single such prospect, reference to the forecasts obtained from one model for tourism policy and investment decision-making could be very risky. Properly combining individual forecasts can reduce this risk.

This study contributes to both the general and tourism forecasting combination literature by providing empirical evidence on the efficiency of combination forecasts. Further research into other advanced combination methods, such as non-linear combination methods, should be considered to evaluate whether they offer improved accuracy over those presented here.

REFERENCES


