https://doi.org/10.1016/j.annals.2014.06.005 This is the Pre-Published Version.

ACCURACY AND BIAS OF EXPERTS' ADJUSTED FORECASTS

Shanshan Vera Lin
School of Hotel and Tourism Management
The Hong Kong Polytechnic University
Hong Kong SAR

Paul Goodwin
School of Management
University of Bath
U. K.

Haiyan Song
School of Hotel and Tourism Management
The Hong Kong Polytechnic University
Hong Kong SAR

The first and third authors would like to acknowledge the financial support of Zhejiang Province Natural Science Foundation (Grant No. Q14G010016) and the Hong Kong Research Grant Council (Grant No. PolyU 5969/13H), respectively.

ACCURACY AND BIAS OF EXPERTS' ADJUSTED FORECASTS

Abstract

This study investigates whether experts' group-based judgmental adjustments to econometric forecasts of tourism demand improve the accuracy of the forecasts and whether the adjusted forecasts are unbiased. The Delphi method was used to aggregate experts' judgmental adjustments and a range of error measures and statistical tests were employed to evaluate forecast accuracy. Regression analysis was used to investigate whether the statistical and judgmentally-adjusted forecasts were unbiased. The hypothesis tests suggested that, on average, the adjustments of the Delphi panel improved forecast accuracy though the group-adjusted forecasts were found to be biased for some of the individual markets. In-depth interviews with the Delphi panellists provided further insights into the biases that were associated with the Delphi surveys.

Keywords: Tourism forecasts, accuracy, bias, judgmental adjustment

1. RESEARCH BACKGROUND

Tourism is a demand-driven, service-oriented industry that is experiencing rapid growth and innovation (Chu, 2008). Along with the phenomenal growth in demand over the past six decades, there has been a corresponding interest in tourism research. Within this context, tourism demand modelling and forecasting has received intensive attention (Song & Li, 2008). Tourism demand studies mainly focus on two aspects: the analysis of the effects of various determinants on demand and the provision of accurate forecasts of future tourism demand. The majority of the published studies on this topic have focused on statistical (time series and econometric) forecasting approaches, with very limited attention being paid to judgmental forecasting approaches in the tourism forecasting literature.

However, it is difficult, if not impossible, to capture such a diverse, dynamic, and changeable phenomenon as tourism demand using the statistical models that incorporate only a limited number of variables (UNWTO & ETC, 2011). Sociological and psychological factors are difficult to express quantitatively, and crises and disasters are impossible to forecast. For their forecasts to be of any practical value, tourism planners and decision-makers must adjust their forecasts and models to deal with a bundle of qualitative factors (Croce & Wöber, 2011). Judgmental inputs to the forecasting process are thus designed to incorporate the knowledge of experts into tourism forecasts in order to improve their quality (Armstrong & Collopy, 1998; Croce & Wöber, 2011). A big challenge in achieving accurate forecasts is to utilize the best aspects of statistical predictions while also exploiting and capitalizing on the value of knowledge and judgmental information which are not taken into account by the statistical forecasts (Armstrong & Collopy, 1998). It would therefore seem to be advantageous to bring these two methods together. The general forecasting literature suggests that combining methods improves forecast accuracy, a finding that holds true for quantitative forecasting,

judgmental forecasting, and the averaging of these two forecasts (Clemen, 1989).

To date, the combination of multiple methods is still not widely accepted as a viable research strategy by academics in the tourism demand forecasting field. However, tourism demand forecasters and practitioners have indicated that such research is necessary to develop and strengthen our understanding of many tourism-related issues. Most tourism forecasting research has been devoted to the area of single-equation modelling approach (i.e. modern econometric models) (Song & Li, 2008; Witt & Witt, 1995), and it is surprising that the considerable advances in judgmental forecasting achieved in other domains have still not received much attention in the tourism forecasting literature. Given the knowledge capital possessed by tourism analysts, the industry could benefit from attempts to exploit this resource to achieve more accurate forecasts.

This study contributes to the tourism forecasting literature by providing empirical evidence on the efficiency of integrating judgmental and statistical forecasts with a particular focus on judgmentally adjusting statistical forecasts using a Web-based forecasting support system designed by the research team of this study. The aims of the study are to build up a systematic framework to integrate judgmental and statistical forecasts in the tourism context which (a) applies econometric forecasting models to generate statistical forecasts; (b) uses a forecasting decision support system, which has never been used in both general and tourism forecasting literature, to structure experts' knowledge and quantify managerial intuition; (c) measures statistical and judgmentally adjusted forecasts using formal measures of accuracy; and (d) explores the reasons for bias. Moreover, this study provides theoretical and practical evidence to further develop a tourism demand forecasting system in support of collaborative forecasting tasks for tourism practitioners, to enhance the system's effectiveness and efficiency, and to improve its forecasting performance. The remainder of this paper is structured as follows: Section 2 reviews the literature on the proposed hypotheses. Section 3 presents the

methodological details in this study. Section 4 summarizes the hypothesis testing results, together with findings from in-depth interviews with the participating tourism experts. Section 5 concludes the study.

2. LITERATURE REVIEW AND HYPOTHESES

Judgmental adjustment of statistical forecasts is a major alternative to combining statistical and judgmental approaches (combining can involve methods ranging from simple means of the component forecasts to Bayesian forecasting). Numerous industry surveys have revealed that judgmentally adjusted statistical forecasting is a common practice. Klassen and Flores (2001) surveyed 117 Canadian firms and found that senior management frequently revised the forecasts. They also found that 80% of the respondents used computer-generated statistical forecasts and then judgmentally adjusted them. Similarly, in a study of 96 corporations in the USA, Sanders and Manrodt (1994) found that 45% of the respondents claimed that they always adjusted statistical forecasts and that 37% did it sometimes. The main reason they gave for revising quantitative forecasts was to incorporate knowledge of the environment. A study by Fildes and Goodwin (2007) found that company forecasters most commonly made adjustments for special events such as advertising and product promotion campaigns and price changes.

A forecaster's goal in judgmentally adjusting a statistical forecast is to improve the quality of the forecasts by combining the relative strengths of statistical and judgmental methods (Armstrong, 2001). The traditional approach to assessing the quality of the forecasts is to measure forecast accuracy (or forecast errors) using one or more measures, such as the root mean squared percentage error (RMSPE) or the mean absolute percentage error (MAPE). In most contexts, accuracy is the top concern in forecasting performance (e.g. Fildes & Goodwin, 2007). Forecast accuracy signifies the level of agreement between the actual values and the forecast values; it is also regarded as the converse of forecast error, which is the difference

between the actual value and the forecast. However, measurements of accuracy do not offer guidance on how to improve forecasts (Musso & Phillips, 2002) and two further properties are also important: bias and efficiency. Tests for bias are intended to check whether forecasts have consistent errors that reflect a systematic tendency for forecasts to be either too high or too low (Smith, Tayman, & Swanson, 2013). Tests for efficiency are intended to check whether forecasts have made optimal use of all available information (Musso & Phillips, 2002, p. 24). This study did not examine the efficiency of the judgmentally adjusted forecasts as the Delphi panellists did not have the information about their past forecast errors when they made their forecasts.

Studies on the accuracy of judgmental adjustments have reported equivocal results. Some researchers have found that judgmental adjustments improve forecast accuracy. Klassen and Flores (2001) found an average improvement in accuracy of 7.2%. Fildes and Goodwin (2007) found a median improvement of about 7% in forecast accuracy measured by the absolute percentage error , which was slightly higher than the results (between 2.6% and 5%) reported in Fildes et al.'s (2006) study.

Some researchers have recommended that caution be exercised when using this adjustment approach because it may harm forecast accuracy. For example, from the results of a controlled experiment that involved the participation of experts and persons with limited training, Carbone et al. (1983) found that judgmental forecasts were significantly less accurate than forecasts generated from statistical methods. Willemain (1989) argued that when statistical forecasts were nearly optimal, adjustment has little impact on accuracy improvement; however, when statistical forecasts are inaccurate, adjustment improves accuracy. In a subsequent study, Willemain (1991) found that judgmental adjustments led to greater accuracy improvement when excess error (calculated from the difference between the errors generated by the Naïve method and the forecasting method in use) is high.

Human judgment is characterized as being associated with a number of biases, such as inconsistency, conservatism, recency, availability, anchoring, illusory correlation, optimism, wishful thinking, and underestimating uncertainty (Lawrence, Goodwin, O'Connor, & Önkal, 2006; Makridakis, Wheelwright, & Hyndman, 1998). One major problem of using judgmental adjustment is that people often read systematic patterns into the noise associated with a time series and this leads to damaging adjustments to statistical forecasts (Lawrence et al., 2006; O'Connor, Remus, & Griggs, 1993). Collopy and Armstrong (1992) contended that judgmental revisions can improve accuracy if forecasters are able to identify the patterns that were missed in the statistical forecasting procedure. Lawrence et al. (2006) suggested that judgmental adjustments should be used to adjust statistical forecasts under two conditions: first, the statistical method is deficient in estimating the underlying patterns of time series; second, the forecaster has access to contextual information that is not included in the statistical method. Sanders and Ritzman (2001) argued that statistical forecasts should be judgmentally adjusted in situations of high uncertainty and where the forecaster has important information that is not available to the statistical method. They suggested that forecasters should make adjustments to compensate for specific events that a statistical model cannot capture or that the time series had not yet included. Research by Wolfe and Flores (1990) and Flores et al. (1992) showed that improvements could be obtained when judgmental adjustments were made to corporate earnings series with high variability. Sanders and Ritzman (2001) also concluded that judgmental adjustments can lead to greater improvements in forecast accuracy when the process is structured, either with a computer-aided decision support system or paper and pencil, rather than being made ad hoc.

In some organizations judgmental adjustments to statistical forecasts are made by groups of forecasters, rather than individuals, at forecast review meetings (Fildes et al., 2009). Groups are likely to have access to a wider range of information than individuals and they can also

bring the benefits of multiple perspectives and debate and discussion to the forecasting process (Lock, 1987). However, these benefits may be lost when particular members of the group dominate discussion or status differences between members inhibit the contributions of more junior people (Lock, 1987). The Delphi method is designed to overcome these problems by allowing individuals to put forward their judgments anonymously. These judgments are then summarised by a facilitator (usually in the form of medians and other statistical measures) and fed back to the group members (or panellists) who are invited to revise their original judgments, if they see fit (Frechtling, 2001; Lock, 1987). The process then proceeds over a number of rounds and usually the median estimate of the group in the last round is used as the forecast (Parenté & Anderson-Parenté, 1987).

Rowe and Wright (1999) found evidence that the Delphi method tends to improve judgments obtained from groups but recommended that panellists should also be encouraged to circulate anonymous written discussion otherwise their fellow panellists will have little basis for changing their judgments between rounds. The Delphi method has a number of other practical advantages. It allows people to change their mind without loss of face and panellists do not need to be in the same geographical location so the method naturally lends itself to implementation on the Internet. However, the benefits of anonymity in the Delphi process come at a cost in that the exchange of information between panellists is restricted and there is almost no opportunity for debate. Nevertheless, a recent experimental study carried out by Song, Gao, and Lin (2013) suggested that integrating statistical and judgmental forecasts in a Web-based forecasting system through a dynamic online Delphi survey could significantly improve forecast accuracy in the tourism context. Their study focused on the design of the Web-based system forecasting support system and carried out a limited experiment using judgmental adjustments by academics and students. The current study, however, is more comprehensive in terms of the forecasting horizon and the composition of the Delphi panel,

which includes both practitioners and academics. In addition to the traditional forecast error evaluation, this study also investigates the bias of the statistical and judgmental forecasts. Moreover, in-depth interviews were conducted to provide qualitative input to interpret the quantitative findings. Based on the findings of Song et al. (2013), hypothesis *H1* was formulated:

H1: Delphi-based judgmental adjustments to statistical forecasts improve the accuracy of tourism forecasts.

The relative accuracy of statistical forecasts compared to those generated by the simplest Naïve 1 model, which sets all forecasts equal to the most recent observation, is of particular interest. In order to be a useful forecasting tool, it is generally accepted that forecasting models should be able to make forecasts that are at least as accurate as those generated by a Naïve no change model. Naïve 1 model is commonly accepted as a useful benchmark for forecasting comparison (Lawrence, O'Connor, & Edmundson, 2000; Lin, 2013; Makridakis et al., 1982; Makridakis et al., 1993; Witt & Witt, 1995), and is therefore considered in this study. Hypothesis *H2* was developed accordingly:

H2: Delphi-based judgmentally adjusted forecasts are more accurate than Naïve forecasts.

While judgmental adjustments may improve the accuracy of forecasts they may still suffer from bias. These may, at least in part, result from the use of heuristics by judgmental forecasters – a heuristic is a simplified mental strategy that people use to cope with the complexity of the forecasting task (Tversky & Kahneman, 1974). For example, the anchor and adjustment heuristic involves making an estimate by identifying a starting value (the anchor) and adjusting from this to make a final estimate. However, there is a tendency for the adjustment from the anchor to be insufficient. Hence, people may anchor on the most recent value in a time series

and under adjust from it when making forecasts. As a result growth or decay in series tends to be systematically underestimated (Eggleton, 1982; Lawrence & Makridakis, 1989; Sanders, 1992; Wagenaar & Sagaria, 1975), leading to bias in the forecasts. Similarly, the use of the availability heuristic means that the probability of future events is assessed on the basis of the number of instances of the event that can be brought to mind. Thus, recent events or events highlighted by the media may be overweighted when compared to less salient events or those that occurred in the more distant past. This heuristic may also lead to illusory correlation where people see pre-conceived relationships between the available information and the variable-to-be forecast that do not exist (Chapman & Chapman, 1969).

Lawrence, O'Connor, and Edmundson (2000) reported that studies of real world judgmental forecasting have all found bias in the forecasts. In a field study of forecasting in 13 manufacturing organizations, they found that these deficiencies were sufficient to outweigh any contribution to accuracy of the contextual information that the managers had access to, but which was not available to the statistical forecasts.

The relationship between accuracy and bias is not necessarily straightforward. Unbiasedness cannot guarantee high accuracy. For example, Ali, Klein, and Rosenfeld (1992) concluded that the accuracy of short-term forecasts in predicting annual earnings per share is not improved through the adjustment procedure, even though the adjustment behaviour leads to reductions in bias and serial correlation. Mathews and Diamantopoulos (1986, 1990) showed that judgmental adjustment could introduce bias even when it improves forecast accuracy. Fildes, Goodwin, Lawrence, and Nikolopoulos (2009) also found that although judgmental adjustments may help to improve accuracy, they can also be either biased or inefficient. Given the findings that bias has been found to be endemic in judgmental forecasting in other domains, we arrive at the following hypothesis:

H3: Delphi-based judgmentally adjusted forecasts of tourism demand are biased.

3. METHODOLOGY

A substantial volume of the research on judgmental adjustments has been conducted in experimental settings, such as psychology laboratories, that may or may not be representative of an actual organizational setting. Goodwin and Wright (1993) identified 11 ways in which a lab-based experimental study might fail to represent an organizational setting where the statistical forecasts were judgmentally adjusted. Therefore, the results of experimental studies relevant to judgmental adjustments may not always be generalizable and researchers have been encouraged to conduct more studies in realistic conditions (Önkal & Gonul, 2005). The Delphi panel in the current study was recruited from different sectors of the tourism and hospitality industry in Hong Kong, including academic institutions, the accommodation sector (e.g. hotels, resorts), tourist attractions/tourist facilities, travel trades (e.g. tour operators, travel agents), and government offices. Unlike experimental studies in which artificial data are often used, the use of actual decision makers in real-world forecasting conditions provides external validation, thus making the findings from this study more convincing and reliable.

A mixed methods approach − the sequential explanatory strategy summarized by Creswell (2009) − was adopted to utilize the combined strengths of qualitative and quantitative approaches. The quantitative techniques comprehensively evaluated the two dimensions of forecasting performance (accuracy and bias). In-depth interviews were then carried out among those experts who participated in the main Delphi surveys a year later to investigate the reasons for the causes of biases in order to attempt to reduce bias in future judgmental adjustments to statistical forecasts. A total of 14 experts (five industry experts and nine academic experts) participated in the interviews.

3.1 Variables and data

The most commonly used variable in measuring international tourism demand is visitor arrivals from an origin country/region to a given destination, followed by tourist expenditure and tourist nights in registered accommodation in the destination (Song & Li, 2008; Song, Lin, Zhang, & Gao, 2010; Song, Witt, & Li, 2009). In this study, the visitor arrivals variable was selected to measure inbound tourism demand in Hong Kong.

According to the existing literature, the most commonly considered influencing factors of tourism demand are the income level of the origin country/region, the tourism price of the destination relative to that of the origin country/region (i.e. the own price of the tourism products), tourism prices in competing destinations (i.e. substitute prices) (the own prices and substitute prices are often adjusted by the relevant exchange rates), and travel costs from the origin countries/regions to the destination (Song, Kim, & Yang, 2010; Song & Li, 2008; Song, Witt, & Li, 2009). However, several empirical studies, for example, Kim and Song (1998), have suggested that the travel cost variable is insignificant in certain tourism demand models. Some studies have also included lagged dependent variables in their regression models. It is also important to note that other factors such as marketing expenditure of the tourism product/service providers (both at the destination and firm level), and the change of tastes and preferences towards Hong Kong as a tourist destination in the source markets can play a role in the determination of tourists travelling to a destination. The difficulty in accessing the relevant marketing data hinders its application in most empirical studies (Kulendran & Dwyer, 2009; Zhang, Kulendran, & Song, 2010). The demand model used here drew on data from a range of publicly available sources. Quarterly data from 1985Q1 to 2010Q4 were used to estimate the demand models, which were then used to generate the quarterly forecasts from 2011Q2 to 2015Q4. The data of the dependent variable, measured by visitor arrivals, were

collected from the *Visitor Arrival Statistics* (HKTB, 2011). This is the best data available for the purposes of the modelling exercise for this study. The GDP data were collected from the *International Financial Statistics* (IFS) of the International Monetary Fund (IMF, 2011) and the official websites of the statistical bureaus or departments of all countries and/or regions concerned. Consumer price indexes (CPIs) (2005=100) and exchange rates were also obtained from the IMF. Six of Hong Kong's competing destinations, Mainland China, South Korea, Japan, Singapore, Taiwan, and Thailand, were selected to calculate the substitute prices. The inbound visitor arrivals of six selected origins (i.e. Mainland China, Japan, Taiwan, Australia, the UK, and the USA) to these six destinations were respectively collected from the official websites of HKTB (2011), Korea Tourism Organization (2011), Japan National Tourist Organization (2011), Singapore Tourism Board (2011), Tourism Bureau Ministry of Transportation and Communication in Taiwan (2011), and Department of Tourism in Thailand (2011).

3.2 The econometric model

In line with the majority of the tourism demand literature such as Chon, Li, Lin and Gao (2010), Song, Kim, and Yang (2010), Song and Lin (2010), Song, Lin, Witt, and Zhang (2011), and Song et al. (2013), the following autoregressive distributed lag -Error correction model (ARDL-ECM) was employed to model and forecast the inbound tourism demand in Hong Kong.

$$\begin{aligned}\Delta \ln VA_{it} = {} & \alpha_0 + \sum_{j=1}^{p_1} \psi_{Qj} \Delta \ln VA_{i,t-j} + \sum_{j=0}^{p_2} \psi_{Yj} \Delta \ln Y_{i,t-j} \\ & + \sum_{j=0}^{p_3} \psi_{Pj} \Delta \ln P_{i,t-j} + \sum_{j=0}^{p_4} \psi_{P_{is},j} \Delta \ln P_{is,t-j} \\ & + \pi_1 \ln VA_{i,t-1} + \pi_2 \ln Y_{i,t-1} + \pi_3 \ln P_{i,t-1} + \pi_4 \ln P_{is,t-1} \\ & + \delta_1 D_1 + \delta_2 D_2 + \delta_3 D_3 + \sum_{d=1}^{D} \theta_d Dummies + u_{it}\end{aligned} \tag{1}$$

where Δ is the first difference operator (i.e. $\Delta X_t = X_t - X_{t-1}$), VA_{it} is the tourism demand variable measured by visitor arrivals from the i^{th} source market to Hong Kong at time *t*, Y_{it} is the income of tourists from the i^{th} source market at time *t*, D_1, D_2, and D_3 were seasonal dummy variables to capture the influence of seasonality in the dependent variable (visitor arrivals), *Dummies* were one-off event dummy variables to capture influences of such events as the 9/11 terrorist attack, SARS, and other destination specific events relevant to the demand for Hong Kong tourism, and ε_{it} is an error term assumed to be normally distributed with zero mean and constant variance, i.e. $u_{it} \sim N(0,\sigma^2)$. The lag order p_i (i = 1, 2, 3, and 4) in Equation (1) was determined by the Akaike Information Criterion (AIC). This study adopts the AIC instead of the Schwarz Bayesian Criterion (SBC) in the lag length selection because "the AIC model appears to be statistically more acceptable than the SBC criterion" (Halicioglu, 2008, p. 8). The own price (P_{it}) is the price of tourism in Hong Kong at time *t* relative to that of the i^{th} source market, and is given as:

$$P_{it} = (CPI_t^{HK} / EX_t^{HK}) / (CPI_t^i / EX_t^i) \quad (2)$$

where CPI_t^{HK} and CPI_t^i are the *CPI*s for Hong Kong and the i^{th} origin country/region at time *t*, respectively, and EX_t^{HK} and EX_t^i are the exchange rate indexes for Hong Kong and i^{th} origin country/region at time *t*, respectively (all exchange rates were calculated based on the local currencies against the US dollar).

The substitute price (P_{ist}) is calculated as a weighted index of *CPI* of each of the six substitute markets according to its share of international visitor arrivals at time *t*, which is given as:

$$P_{ist} = \sum_{j=1}^{6} (CPI_{jt} / EX_{jt}) w_{jt}^i \quad (3)$$

where j = 1, 2, 3, 4, 5, and 6, representing China, South Korea, Japan, Singapore, Thailand and Taiwan, respectively, w_{jt}^{i} is calculated as $TVA_{jt}^{i} / (\sum_{j=1}^{6} TVA_{jt}^{i})$, indicating the share of international visitor arrivals for the j^{th} country/region at time t, and TVA_{jt}^{i} is the visitor arrivals of substitute destination j from origin country/region i at time t.

The reason for selecting the ARDL-ECM was made on two grounds. First, the forecasts generated by ARDL-ECM were found to be highly accurate as the average MAPEs of the forecasts were around 5% based on an annual evaluation of accuracy over 2010−2012 (HKTDFS, 2011, 2012). Second, the modelling and forecasting procedure of ARDL-ECM was embedded in the Hong Kong Tourism Demand Forecasting System where the Delphi survey and integration were carried out.

3.3 Judgmental adjustment

The integration of statistical and judgmental forecasting in this study was defined as the *voluntary integration* of statistical forecasts with Delphi panellists' group judgment rather than the *mechanical integration* of two forecasts. Voluntary integration, as described by Goodwin (2000), is the process of supplying judgmental forecasters with statistical forecasts that they can ignore, accept, or adjust. In this study, the Hong Kong Tourism Demand Forecasting System (HKTDFS) was applied to produce the voluntary integration of statistical forecasts and Delphi experts' judgmental inputs. A more detailed introduction of HKTDFS and methodological details can be found in Song et al. (2013).

The final panel consisted of 11 academic researchers (61%) and seven industry practitioners (39%). Over half (58%) of the panellists who were contacted responded to the Delphi survey in the first round; a lower positive return rate (54.8%) was achieved in the second round. In total, 15 experts took both the first (17 June to 6 July 2011) and second round (11

July to 27 July 2011) surveys. Panellists were invited to make their adjustments to the econometric forecasts of visitor arrivals to Hong Kong from three short-haul markets (i.e. China, Taiwan, and Japan) and three long-haul markets (i.e. the USA, the UK, and Australia) over 2011Q2−2015Q4. This survey considered the impact of the Japanese earthquake in 2011, the construction of a high-speed railway between China and Hong Kong, the London Olympic Games in 2012, and the opening of three new themed lands in the Hong Kong Disneyland.

3.4 Evaluation of forecast accuracy

All forecast error measures have limitations and the relative accuracy of forecasting methods may vary depending on which measure is used. Because of this a range of error measures were selected to evaluate the performance and accuracy of the forecasts in this study: the percentage better (PB) (than comparison forecasts), absolute percentage error (APE), mean absolute percentage error (MAPE), root mean squared percentage error (RMSPE), and Theil's U statistic (U statistic). A smaller value of all of the measures (except for the U statistic) indicates greater accuracy. The advantage of using the U statistic lies in the fact that it "allows a relative comparison of formal forecasting methods with Naïve approaches and also squares the errors involved so that large errors are given much more weight than small errors" (Makridakis, Wheelwright, & Hyndman, 1998, p. 48).

Lewis (1982) has suggested that if the MAPE of a model is less than 10%, it is a highly accurate forecasting model, but much depends on the context and how predictable the forecast variable is. Both the MAPE and RMSPE allows comparison of accuracy across time series measured on different scales, but tend to be distorted when actual values are low. In addition to the conventional measures of forecast accuracy, the PB, which counts and reports the percentage of time that a given forecast has a smaller forecast error than another forecast, was also used to evaluate forecast accuracy in this study.

3.5 Tests for the bias of judgmental forecasts

The studies by Ali, Klein, and Rosenfeld (1992), Harris (1999), and Lawrence, O'Connor, and Edmundson (2000), indicate that the bias of judgmental forecasts can be investigated by fitting a regression model using the following equation:

$$PE_t = \alpha_0 + \beta_0 PE_{t-1} + \mu_t \qquad (4)$$

where $PE_t = (A_t - F_t) / A_t$, A_t is the actual value at time period t, and F_t is the forecast made for period t.

Bias is defined as "the average difference between the actual value of each variable and its forecast value" (Batchelor, 2001, p. 228). In other words, if there is no bias in the forecasts, α_0 is expected to be zero. If there is a consistent pattern of underforecasting (or overforecasting), α_0 should be positive (or negative). A negative α_0 coefficient means that the average forecast error is less than zero, suggesting that there is a consistent pattern of overforecasting (Harris, 1999; Lawrence, O'Connor, & Edmundson, 2000). A positive α_0 coefficient shows that the average forecast error is greater than zero, indicating that there is a consistent pattern of underforecasting. The rejection of the null hypothesis that α equals zero shows that, on average, experts' forecasts display a level bias.

As an alternative test of forecast biases, the percentage of cases where the arrivals forecast was greater than the actual figures was calculated for each round and the binomial test was used to determine whether this was significantly different from the 50% figure that is expected in unbiased forecasts. If forecasts are unbiased, the frequency of underforecasts (or positive forecast errors) should, on average, be the same as that of overforecasts (or negative forecast errors).

4. FINDINGS AND DISCUSSIONS

The first section presents the results of the hypothesis testing − it provides an extensive analysis of the biasness and accuracy of the statistical and judgmental forecasts. In addition to the supporting evidence from the literature, the findings from in-depth interviews in the second section aimed to investigate the causes of the biases in the judgmental adjustments from the experts' viewpoint.

4.1 Hypotheses testing results

Forecast accuracy was evaluated by comparing the MAPE and RMSPE of the forecasts generated by the econometric model against the forecasts that were judgmentally adjusted by the Delphi panellists. Associated statistical tests were carried out to examine whether there was any significant difference between the two groups of forecasts.

As shown in Table 1, the judgmentally adjusted forecasts were more accurate than the statistical forecasts alone (i.e. baseline forecasts) when assessing the accuracy from 2011Q2−2012Q2: the mean MAPE decreased from 8.6% to 7.5% in the initial round (R1) and to 6.5% in the subsequent round (R2). The percentage reductions of MAPE ranged from 9.0% to 24.6%, and even larger reductions were found in RMSPE (from 17.8% to 36.9%). After the experts' judgmental adjustments, none of the MAPEs exceeded 20%, suggesting that post-adjusted forecasts were more satisfactory than those forecasts without adjustments. This finding was consistent irrespective of whether the MAPE or RMSPE was used to evaluate the forecast accuracy. Table 1 shows that the results obtained by RMSPE were globally similar to the ones obtained with MAPE; this finding held true for individual forecasting horizons (quarters).

[Insert Table 1 here]

Wilcoxon signed-rank tests were applied to examine if any significant difference existed between the accuracy of statistical forecasts, Round 1, and Round 2 forecasts using the APE as the accuracy measure. The results in Table 2 show that the average of the Delphi group's Round 1 forecasts did not significantly outperform the statistical forecasts (Z = -0.46, p = 0.33; T =17, r = -0.06). However, the Round 2 forecasts were significantly more accurate (at least the 10% level) than both the statistical forecasts and the Round 1 forecasts − the p values were 0.09 and 0.004 respectively − indicating the benefits of the conducting multiple rounds in the Delphi process.

[Insert Table 2 here]

Not only did the forecast adjustments improve the overall forecast accuracy, the improvements were also evident across markets and over different rounds of Delphi (see Table 1). Tables 1-2 suggest that the largest accuracy improvement over the statistical forecasts was found in the prediction of visitor arrivals from China, followed by Japan, and the least improvement in accuracy over the statistical forecasts was found in the prediction of visitor arrivals from Australia. The relatively poor performance of the experts' adjustments for Australia and the USA could be attributed to the already good performance of the statistical forecasts (below 3%). When similar comparisons to those shown in Table 2 were made using APE, the results were found to be similar in most cases.

Table 3 provides a more detailed analysis of the performance statistics for individual quarters by markets and rounds as assessed by APE. The APEs of the three sets of forecasts (SF, GF1, and GF2) were calculated for each quarter between 2011Q2 and 2012Q2. The percentage of times that adjustments reduced the APE for the five forecasts in each market are shown in Table 3, together with the percentage of times that the Delphi Round 2 forecasts improved on those in Round 1. Reductions in the APE or an improvement in forecast accuracy

as a result of using the forecasting adjustment method (versus statistical forecasting alone) were observed in five of the six markets (the exception being the UK) in Round 1 and in all six markets in Round 2. However, improvements in APE varied across different markets. Similar to the findings obtained from Song et al. (2013), this confirmed that forecasts for the long-haul markets were more accurate than those for the short-haul markets.

[Insert Table 3 here]

With regard to the Mainland market, there was an improvement in forecast accuracy after judgmental adjustment either in Round 1 or Round 2 as the PB statistics show that error reductions of APE were found in all quarters. For the UK market, forecasting adjustment only produced accuracy improvement in Round 2. For the Taiwan and Japan markets, the accuracy of forecasts was improved after adjustment and was particularly evident in the final round. As measured by APE, forecast accuracy in terms of predicting the number of Japanese visitors ranged from 1.9% to 19.3%. This was probably due to the impact of the earthquake in March 2011, which not only seriously affected the quarter of the year in which the disaster happened but also the subsequent year. For the USA market, although accuracy improved in Round 1, the improvement decreased with iteration as the PB statistic reduced from 80% to 20%.

In short, the above analysis shows that, on average, judgmental revisions of the statistical forecasts led to an improved accuracy in predicting visitor arrivals to Hong Kong which was particularly true after iteration. The above findings support hypothesis *H1*.

As a benchmark against which to compare the accuracy of the experts' judgmentally adjusted forecasts and the statistical forecasts, the performance of forecasts made by the Naïve 1 method were considered by calculating the U statistic. The overall performance of the statistical forecasts was similar to that of the Naïve 1 forecasts in predicting Hong Kong inbound tourism flows as the U statistic was 1.03, marginally larger than unity. The U statistic

of the statistical forecasts for short-haul markets (1.64) was much higher than that of the long-haul markets (0.29). After adjustments, the U statistics reduced from 1.20 (Round 1) to 1.02 (Round 2) for the short-haul markets. For the long-haul markets, the U statistics were also observed to decrease from 0.35 in Round 1 to 0.30 in Round 2, which was higher than the value for the initial statistical forecasts (0.29). The above findings backed up the hypothesis *H2* that, on average, judgmentally adjusted forecasts are more accurate than Naïve forecasts.

We should be cautious in interpreting this finding as the value of the U statistic could have been determined by the accuracy of two factors: the inclusion of six source markets with different degrees of forecasting difficulty, and a mix of multiple-step forecasts. A further examination of the U statistic results by markets in Table 1 shows that the high value of the U statistic was mainly due to the Mainland market, which had a relatively large value (2.63 for SF, 1.82 for GF1, and 1.53 for GF2). The other five markets all had U statistics below one, which suggests that the adjusted and unadjusted forecasts for these five markets were, on average, better than the Naïve forecasts.

The literature shows that forecasts produced by models are generally better than unaided judgment, except where special circumstances apply, and that the use of judgment can introduce biases (Stekler, 2007). People's predictions are therefore likely to contain at least some component of systematic errors (Armor & Taylor, 2002). It is therefore important to investigate the extent to which the experts' adjustments were biased here.

To test for the bias of the judgmentally adjusted forecasts, a pooled regression model of Equation (4) was estimated over the sample period 2011Q2 to 2012Q2. The statistical analysis of forecast errors was based on the null hypothesis of no bias. Table 4 reports the results of the regression analysis clustered by source market. The first pooled regression model was estimated by using the group forecasts − the average of individual forecasts in each round, namely G1

and G2. The results suggest that the adjusted forecasts for Round 1 and Round 2 were unbiased: α was insignificant, and hence there was no evidence of bias in the forecasts either in the first or second round.

[Insert Table 4 here]

To investigate whether or not the adjusted forecasts made by individual experts were biased, Equation (4) was re-estimated using the pooled sample of all of the individual experts' adjusted forecasts in each round. It was found that the intercept (or constant) for the second pooled regression model was not significantly different from zero which again provided no evidence that the individual experts' adjusted forecasts were biased.

In addition to the regression analysis, an alternative test of forecast bias − the percentage of cases where the forecast (either adjusted or unadjusted) was greater than the actual value was computed and the binomial test was used to determine whether this was significantly different from 0.5 (50%). The binomial tests shown in Table 5 confirmed the results from the regression analysis, which showed that the statistical forecasts and group forecasts in Round 1 and Round 2 were, on average, unbiased as the *p* values for the three sets of tests were all above 0.05.

[Insert Table 5 here]

Even though the regression analysis and the binomial test results both suggested that the three sets of forecasts − the statistical forecasts and the group forecasts in Rounds 1 and 2 − were unbiased, it is necessary to be cautious about concluding that arrival forecasts from all of Hong Kong's source markets are unbiased. Instead, it is more reasonable to assume that different biases in different series cancelled each other out, as suggested by Harvey (2007). An examination of the historical arrivals trends of the six source markets found that there was a mixture of different trends that could possibly cancel out the biases from individual markets. For example, the growth of the arrival series for the Mainland market appeared to be

exponential, while the trend for the Japan market has remained quite stable in the past 3 decades. It is thus valuable to not only investigate all forecasts (with a mixed structure of different arrival trends) but also forecasts from individual markets, which will help us to gain a better understanding of whether the final forecasts were truly unbiased or not.

A closer analysis of the individual market results revealed that the majority of the forecasts overestimated future arrivals. Figure 1 provides visual evidence of the direction of the bias for individual markets. It can be seen from Figure 1 that forecasts from Australia, Taiwan, and the UK were overestimated while forecasts from China were underestimated. In terms of the Japan market, the experts' forecasts were too optimistic in evaluating the impacts of the earthquake of March 2011 on Hong Kong's inbound tourism industry. It seems that there was an overforecasting tendency in estimating the number of Japanese visitors to Hong Kong in the second quarter forecasts over the forecasting period 2011 to 2015.

[Insert Figure 1 here]

The line plots only provided visual information; the regression analysis gave further information to confirm the bias tests among individual markets. The negative intercept terms in Table 4 suggest that the Delphi experts, on average, overestimated visitor arrivals, although this was not found to be statistically significant. For individual markets, it was also found that the intercept term was significantly different from zero for four of the six markets (Australia, China, Taiwan, and the UK) in Round 1 and five of the six markets (Australia, China, Taiwan, the UK, and USA) in Round 2.

Generally, the experts consistently overestimated visitor arrivals for all of the markets except for China. One explanation for the tendency to underforecasting in the Mainland market is probably the incredibly increasing growth trend in this market in the past 3 decades. This is consistent with previous studies, such as, Wagenaar and Sagaria (1975), Eggleton (1982),

Lawrence and Makridakis (1989), and Sanders (1992), that have suggested that people appear to underestimate the steepness of trends in series and tend to underestimate upward trends. In a more recent study, Harvey (2007) also found that forecasts from linear and exponential trends would show underadjustment. Critics have noted that judgments tend to be too conservative in the face of rapid change, typically underestimating exponential growth. In Mathews and Diamantopoulos's (1989) study in a health products company, the evidence of an optimism bias in managers' revisions of forecasts was found. They explained that these adjustments may have been partly a reaction to systematic underestimation by statistical forecasting models.

The tendency for overforecasting in most adjustments may be explained by the existence of optimism bias. As noted by Armor and Taylor (2002), one of the most robust findings in the psychology of prediction is that people's predictions tend to be optimistically biased. According to one of the leading explanations for why people exhibit optimistic biases, people tend to "infer the likelihood of different outcomes on the basis of case-specific plans or scenarios about how the future will unfold" and "the very processes of constructing and considering these scenarios tend to render people prone to bias" (Armor & Taylor, 2002, p. 342) in that they lead to attention being paid to the case-specific factors rather than the underlying base-rate which may suggest that a less optimistic forecast should be made.

To sum up the findings, the evidence presented in this section suggests that judgmentally revised forecasts were, on average, unbiased. Thus, the findings support *H3*. Given that experts' predictions are biased, their forecasting performance should be monitored based on the history of their interaction with the system. During the judgmental forecasting process, they should be alerted against any systematic bias.

4.2 In-depth interviews: The role of heuristics

In-depth interviews with the experts were used to obtain further insights into the process

they used to produce their adjusted forecasts. One possibility was that the experts anchored on the most recent data point so that any adjustments they made to take into account non-time series information that they possessed were too small. This would be possible because the experts were provided with a full view of the historical data (in graphical format) and recent 4-year data (in graphical and tabular format) from which to make adjustments in the HKTDFS. The majority of the respondents reported that they checked the historical trends of visitor arrivals and considered them in their adjustment process. One academic expert explained that he believed that the historical trend/pattern of an arrival series is a good indicator in terms of projecting the future. If the experts were accurately reporting how they arrived at their adjustment then this would not be consistent with a tendency to anchor on the most recent data point, instead it would suggest that the long term patterns was being used to guide the adjustments. For example, the following quote shows how one academic expert made his adjustments:

> *When doing the adjustments, I try to make it not diverge from historical trend too much. I remember some forecasts tend to be far away from the recent years. I have to adjust it to make it a little bit close to the normal trend.*

However, anchoring on the last data point and making insufficient adjustment from it has been used to explain the tendency for judgmental forecasts to "[be] below the optimum for upward trends but above the optimum for downward ones" (Harvey, 2007, p. 17). Two experienced industry respondents observed that in the past few years, they had been too conservative in forecasting tourism demand in Hong Kong, particularly for the Mainland market. Although they acknowledged the massive growth of the demand in their forecasting assumptions, their forecasts had still always been lower than the actual arrival figures. Thus the evidence for the use of anchoring and adjustment is contradictory. This may be due to the difficulty that experts had in accurately recalling and reporting how the actually arrived at their

adjustments. Making the adjustment itself may be an intuitive process, invoking rapid, unconscious, and system 1 cognitive processing (Kahneman, 2012), while explaining the process of adjustment will involve explicit, conscious and deliberative (or system 2) processing.

In terms of the statistical forecasts provided, most of the experts took them as the continuation of the historic data, and thus the majority of them said that they checked both the historical data and the statistical forecasts in order to make reasonable adjustments. There was a strong majority who stated that they only adjusted the statistical forecasts if they believed that it was absolutely necessary. This is consistent with the observed behaviour of the experts and is an example of good forecasting practice. Usually, there is a tendency for people to adjust statistical forecasts too frequently partly as a result of the use of the representativeness heuristic (Fildes et al., 2009). If this heuristic was being employed it might have caused the experts to gratuitously adjust the statistical forecasts so that they had an irregular pattern that they judged to be representative of the noisy pattern observed in the past, rather than adjusting for the effect of special events. One reason for this good practice may have been the request for the experts to provide explicit reasons for their adjustments during the Delphi process.

The use of the availability heuristic suggests that people tend to consider information that is more easily retrieved from memory to be associated with more likely events (Harvey, 2007). In the interviews, the academic experts stated that they usually considered factors such as economic conditions, historical trends, and the impact of ad hoc events in the forecasting period; for example, one expert claimed that she relied on historical trends to compare the provided statistical forecasts and incorporated the effects of special events (e.g. the Japan earthquake, 2012 London Olympic Games) by making adjustments to the initial forecasts. One possible bias caused by the availability heuristic is “recency”. Makridakis, Wheelwright, and Hyndman (1998) described “recency” as a type of bias in which the most recent events dominate those in the less recent past, which are downgraded or ignored. This type of bias was

also found among a few of the interviewed experts. One academic interviewee revealed that her adjustments were largely made through recalling recent events; for example, when adjusting forecasts from the Japan market, she considered the impact of the earthquake of March 2011.

5. CONCLUSIONS

This study employed a group of error measures to examine forecast accuracy of statistical forecasts of tourism demand both before and after they had been adjusted by a Delphi panel of experts. In addition, the potential bias of the forecasts was investigated by regression analysis and statistical tests and the causes of bias were further explored by analyzing in-depth interviews results. A traditional comparison to a Naïve forecast was also made. Of course, the results are based on events occurring in the five quarters for which the forecasts are made though they also relate a wide range of markets. If these periods were in some way unusual then the results may not generalize, but we have no reason to think that they were untypical.

The effectiveness of judgmental adjustments is evaluated by examining the accuracy of judgmentally adjusted forecasts compared to the initial statistical forecasts. In this study, the results of the hypothesis tests showed that, on average, the judgmental adjustments made on the basis of statistical forecasts improved accuracy, particularly after an iteration in the Delphi process. This finding was consistent across a range of accuracy measures. Not only did the forecast adjustments improve the overall forecast accuracy, the improvements were also evident across markets. Improvements in accuracy over the initial statistical forecasts were observed in the consensus group forecasts in Rounds 1 and 2 for all of the six source markets. As a judgmental method, the Delphi group forecasting technique is prone to human bias, although structured procedures help to control this. The use of the Delphi technique to structure and aggregate experts' adjustments may help to increase the efficiency of the adjusted forecasts

but not to remove bias. The results from testing hypothesis *H3* show that although the consensus group forecasts were, on average unbiased, the experts' adjustments were biased for some individual source markets. It was found that the experts had different tendencies in forecasting different markets. Generally, the Delphi experts in this study tended to be optimistic in their forecasting tasks. The results from the regression analysis show that these experts made more optimistic forecasts than pessimistic forecasts.

The findings from the in-depth interviews identified a few types of bias that were consistent with the results obtained from the main Delphi surveys. The use of different heuristics can produce different biases, such as anchoring and recency. The findings from the interviews provided evidence that the experts had a high reliance on baseline forecasts. To avoid or reduce the negative impact of anchoring bias, it may be useful to ask experts to discuss and quantify the impacts of possible forthcoming events along with the reasons why such events are proposed. To increase the forecast accuracy, it is important to use advanced econometric forecasting techniques such as ECM, time varying parameter (TVP) model, almost ideal demand systems (AIDS) and their variants as they were found to be superior in terms of forecasting accuracy (Goh & Law, 2011).

Given that judgmentally adjusted forecasts are biased for individual markets, it is suggested that some internal debiasing mechanisms should be incorporated into the HKTDFS to help its users at every stage of judgmental adjustment such as selection of baseline forecasts, and provision of feedback. Since no studies testing the bias of judgmental forecasts have been carried out with tourism demand data, the findings from this study provide a valuable starting point for investigating the reasons for forecasting failure and making suggestions to improve forecast accuracy.

The findings showed that the forecasts after experts' adjustments did not always lead to

accuracy improvement, particularly for those matured long-haul markets — experts' judgmental adjustments harmed forecast accuracy for predicting the arrivals from the USA as the initial statistical forecasts were already highly accurate. Under such a condition, judgmental interventions by tourism forecasters are unlikely to significantly improve forecast accuracy; on the contrary, they would probably have a detrimental effect on the accuracy. A mechanical integration of two sets of independent forecasts, statistical and judgmental forecasts, can utilize the best aspects of both methods while reducing bias (Sanders & Ritzman, 2001), and is thus recommended by most researchers (Goodwin, 2000) as compared to the voluntary integration. All the above aspects suggest possible directions for future research.

REFERENCES

Ali, A., Klein, A., & Rosenfeld, J. (1992). Analysts use of information about permanent and transitory earnings components in forecasting annual EPS. *The Accounting Review, 67* (1), 183–198.

Armor, D. A., & Taylor, S. E. (2002). When predictions fail: The dilemma of unrealistic optimism. In T. Gilovich, D. Griffin, & D. Kahneman, *Heuristics and biases: The psychology of intuitive judgement* (pp. 334–347). Cambridge, UK: Cambridge University Press.

Armstrong, J. (2001). Combining forecasts. In J. Armstrong, *Principles of forecasting: A handbook for researchers and practitioners* (pp. 417–440). Dordrecht: Kluwer Academic.

Armstrong, J., & Collopy, F. (1998). Integration of statistical methods and judgment for time series forecasting: Principles from empirical research. In G. Wright, & P. Goodwin, *Forecasting with judgment* (pp. 269–293). Wiley.

Batchelor, R. (2001). How useful are the forecasts of intergovernmental agencies? The IMF and OECD versus the consensus. *Applied Economics, 33*(2), 225–235.

Carbone, R., Andersen, A., Corriveau, Y., & Corson, P. P. (1983). Comparing for different time series methods the value of technical expertise individualized analysis, and judgmental adjustment. *Management Science, 29* (5), 559–566.

Chapman, L. J. & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, 74, 271–280.

Chon, K., Li, G., Lin, S., & Gao, Z. (2010). Recovery of tourism demand in Hong Kong from the global financial and economic crisis. *Journal of China Tourism Research, 6* (3), 259–278.

Chu, F. L. (2008). Analyzing and forecasting tourism demand with ARAR algorithm. *Tourism Management, 29* (6), 1185–1196.

Clemen, R. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5* (4), 559–583.

Collopy, F., & Armstrong, J. S. (1992). Expert opinions about extrapolation and the mystery of the overlooked discontinuities. *International Journal of Forecasting, 8*, 575–582.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage.

Croce, V., & Wöber, K. (2011). Judgemental forecasting support systems in tourism. *Tourism Economics, 17* (4), 709–724.

Department of Tourism (Thailand). (2011). *Tourist arrivals in Thailand.* Retrieved April 8, 2011, from Department of Tourism (Thailand): http://tourism.go.th/2010/en/statistic/tourism.php

Eggleton, I. R. (1982). Intuitive time-series extrapolation. *Journal of Accounting Research, 20* (1), 68–102.

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces, 37* (6), 570–576.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting, 25* (1), 3–23.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2006). *Producing efficient demand forecasts.* Lancaster University Working Paper.

Flores, B. E., Olson, D. L., & Wolfe, C. (1992). Judgmental adjustment of forecasts: A comparison of methods. *International Journal of Forecasting, 7* (4), 421–433.

Frechtling, D. C. (2001). Forecasting tourism demand: Methods and strategies. Oxford: Butterworth-Heinemann.
Goodwin, P. (2000). Correct or combine? Mechanically integrating judgmental forecasts with statistical methods. *International Journal of Forecasting, 16* (2), 261–275.
Goodwin, P., & Wright, G. (1993). Improving judgmental time series forecasting: A review of the guidance provided by research. *International Journal of Forecasting, 9* (2), 147–161.
Halicioglu, F. (2008). *An econometric analysis of aggregate outbound tourism demand of Turkey.* Retrieved September 2, 2012, from Munich Personal RePEc Archive (MPRA): http://mpra.ub.uni-muenchen.de/6765/
Harris, R. (1999). The accuracy, bias and efficiency of analysts' long run earnings growth rates. *Journal of Business Finance & Accounting, 26* (5–6), 725–755.
Harvey, N. (2007). Use of heuristics: Insights from forecasting research. *Thinking & Reasoning, 13* (1), 5–24.
HKTB. (2011). *Visitor arrival statistics (1985-2011).* Retrieved April 3, 2011, from Hong Kong Tourism Board: http://partnernet.hktb.com/
HKTDFS. (2011). Forecast accuracy of tourist arrivals. Retrieved August 10, 2011, from Hong Kong Tourism Demand Forecasting System (HKTDFS): http://www.tourismforecasting.net/hktdfs/
HKTDFS. (2012). The Hong Kong Tourism Forecasting Reports (2010-2012). Retrieved June 20, 2012, from Hong Kong Tourism Demand Forecasting System (HKTDFS): http://www.tourismforecasting.net/hktdfs/home/project/document.jsp
IMF. (2011). *International Financial Statistics Yearbook*. Retrieved April 25, 2011, from International Monetary Fund: http://195.145.59.167/ISAPI/LogIn.dll/login?lg=e
Japan National Tourist Organization. (2011). *Statistics of visitors to Japan from overseas*. Retrieved September 11, 2011, from http://www.tourism.jp/english/statistics/inbound.php
Kahneman, D. (2011). *Thinking fast and slow.* New York: Farrar, Straus and Giroux.
Kim, S., & Song, H. (1998). Empirical analysis of demand for Korean tourism: A cointegration and error correction and time series models. *International Journal of Forecasting, 13,* 319–327.
Klassen, R. D., & Flores, B. E. (2001). Forecasting practices of Canadian firms: Survey results and comparisons. *International Journal of Production Economics, 70* (2), 163–174.
Korea Tourism Organization. (2011). *Monthly Statistics of Tourism*. Retrieved September 11, 2011, from Key facts on tourism: http://kto.visitkorea.or.kr/enu/ek/ek_4_5_1_2_4.jsp
Kulendran, N., & Dwyer, L. (2009). Measuring the return from Australian tourism marketing expenditure. *Journal of Travel Research, 47,* 275–284.
Lawrence, M., & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior and Human Decision Processes, 43* (2), 172–187.
Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25years. *International Journal of Forecasting, 22* (3), 493–518.
Lawrence, M., O'Connor, M., & Edmundson, B. (2000). A field study of sales forecasting accuracy and processes. *European Journal of Operational Research, 122* (1), 151–160.
Lewis, C. (1982). *Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting.* London: Butterworth Scientific.
Lim, J., & O'Connor, M. (1995). Judgmental adjustment of initial forecasts: Its effectiveness and biases. *Journal of Behavioral Decision Making, 8* (3), 149–168.
Lim, J., & O'Connor, M. (1996). Judgmental forecasting with interactive forecasting support systems. *Decision Support Systems, 16* (4), 339–357.
Lin, S. (2013). Improving forecasting accuracy by combining statistical and judgmental forecasts in tourism. *Journal of China Tourism Research,* 9(3), 325–352.

Lock, A. (1987). Integrating group judgments in subjective forecasts. In G. Wright, & P. Ayton, *Judgmental forecasting* (pp. 109–127). New York: John Wiley & Sons.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., et al. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting, 1* (2), 111–153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The M2-competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting, 9* (1), 5–22.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. (1998). *Forecasting: Methods and applications* (3rd ed.). New York: Wiley.

Mathews, B. P., & Diamantopoulos, A. (1986). Managerial intervention in forecasting. An empirical investigation of forecast manipulation. *International Journal of Research in Marketing, 3* (1), 3–10.

Mathews, B. P., & Diamantopoulos, A. (1989). Judgmental revision of sales forecasts: A longitudinal extension. *Journal of Forecasting, 8* (2), 129–140.

Mathews, B. P., & Diamantopoulos, A. (1990). Judgmental revision of sales forecasts: Effectiveness of forecast selection. *Journal of Forecasting, 9* (4), 407–415.

Musso, A., & Phillips, S. (2002). *Comparing projections and outcomes of IMF-supported programs.* IMF Staff Papers (Vol. 49, No. 1), International Monetary Fund.

O'Connor, M., Remus, W., & Griggs, K. (1993). Judgmental forecasting in times of change. *Journal of Forecasting, 9* (2), 163–172.

Önkal, D., & Gonul, M. S. (2005). Judgmental adjustment: A challenge for providers and users of forecasts. *International Journal of Applied Forecasting, 1*, 13–17.

Rowe, G. & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting, 15*, 353–375.

Sanders, N. (1992). Accuracy of judgmental forecasts: A comparison. *Omega-International Journal of Management Science, 20* (3), 353–364.

Sanders, N., & Manrodt, K. B. (1994). Forecasting practices in US corporations: Survey results. *Interfaces, 24* (2), 92–100.

Sanders, N., & Ritzman, L. (2001). Judgmental adjustment of statistical forecasts. In J. Armstrong, *Principles of forecasting: A handbook for researchers and practitioners* (pp. 405–416). Dordrecht: Kluwer Academic.

Singapore Tourism Board. (2011). *Monthly tourism focus*. Retrieved September 10, 2011, from Tourism statistics publications: https://app.stb.gov.sg/asp/tou/tou03.asp

Smith, S., Tayman, J., & Swanson, D. (2013). Forecast accuracy and bias. In S. Smith, J. Tayman, & D. Swanson, *A practitioner's guide to state and local population projections* (pp. 323–371). Netherlands: Springer.

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.

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., Kim, J. H., & Yang, S. (2010). Confidence intervals for tourism demand elasticity. *Annals of Tourism Research, 37*, 377–396.

Song, H., Lin, S., Witt, S., & Zhang, X. (2011). Impact of financial/economic crisis on demand for hotel rooms in Hong Kong. *Tourism Management, 32* (1), 172–186.

Song, H., Lin, S., Zhang, X., & Gao, Z. (2010). Global financial/economic crisis and tourist arrival forecasts for Hong Kong. *Asia Pacific Journal of Tourism Research, 15* (2), 223–242.

Song, H., Witt, S. F., & Li, G. (2009). *The advanced econometrics of tourism demand.* New York: Routledge.

Stekler, H. (2007). *The future of macroeconomic forecasting: Understanding the forecasting process.* Retrieved August 20, 2012, from http://www.uni-leipzig.de/: http://www.uni-leipzig.de/~forecast/englisch/referenten/Stekler.PDF

Tourism Bureau Ministry of Transportation and Communication (Taiwan). (2011). *Visitor arrivals by residence.* Retrieved April 5, 2011, from Executive Information System: http://admin.taiwan.net.tw/statistics/month_en.aspx?no=14

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185* (4157), 1124–1131.

Tversky, A., & Kahneman, D. (2002). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. In T. Gilovich, D. Griffin, & D. Kahneman, *Heuristics and biases: The psychology of intuitive judgment* (pp. 19–48). Cambridge: Cambridge University Press.

UNWTO and ETC. (2011). *Handbook on tourism forecasting methodologies.* Madrid: World Tourism Organization (UNWTO) and the European Travel Commission (ETC). Retrieved September 20, 2012, from http://www.sete.gr/files/Media/Ebook/2008/110315_Handbook%20on%20Tourism%20Forecasting%20Methodologies.pdf

Wagenaar, W., & Sagaria, S. (1975). Misperception of exponential growth. *Perception and Psychophysics, 18*, 416–422.

Willemain, T. R. (1989). Graphical adjustment of statistical forecasts. *International Journal of Forecasting, 5* (2), 179–185.

Willemain, T. R. (1991). The effect of graphical adjustment on forecast accuracy. *International Journal of Forecasting, 7* (2), 151–154.

Witt, S., & Witt, C. A. (1995). Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting, 11* (3), 447–475.

Wolfe, C., & Flores, B. (1990). Judgmental adjustment of earnings forecasts. *Journal of Forecasting, 9* (4), 389–405.

Zhang, H. Q., Kulendran, N., & Song, H. (2010). Measuring returns on Hong Kong's tourism marketing expenditure. *Tourism Economics, 16* (4), 853–865.

Table 1. Overall Forecasting Performance 2011Q2−2012Q2

Group	Round	MAPE (%)	RMSPE (%)	U
All	SF	8.59	13.04	1.03
	GF1	7.54	10.06	0.79
	GF2	6.47	8.56	0.67
Australia	SF	2.15	2.74	0.33
	GF1	4.03	4.32	0.51
	GF2	3.38	3.57	0.43
UK	SF	5.53	5.83	0.36
	GF1	6.37	6.63	0.40
	GF2	5.24	5.57	0.34
USA	SF	2.07	2.21	0.15
	GF1	1.41	1.85	0.11
	GF2	2.22	2.36	0.16
China	SF	28.18	28.78	2.63
	GF1	19.04	19.98	1.82
	GF2	15.69	16.83	1.53
Japan	SF	8.71	10.65	0.68
	GF1	8.39	9.74	0.63
	GF2	7.26	8.47	0.55
Taiwan	SF	4.89	5.66	0.73
	GF1	6.00	6.87	0.88
	GF2	5.05	5.91	0.76
% Change				
All	GF1-SF	-12.2	-22.9	-23.6
	GF2-SF	-24.6	-34.4	-34.9
	GF2-GF1	-14.1	-15.0	-14.8
Australia	GF1-SF	87.25	57.98	56.27
	GF2-SF	57.07	30.64	31.59
	GF2-GF1	-16.11	-17.31	-15.79
UK	GF1-SF	15.27	13.82	13.09
	GF2-SF	-5.09	-4.45	-4.26
	GF2-GF1	-17.67	-16.05	-15.34
USA	GF1-SF	-31.84	-16.33	-23.92
	GF2-SF	7.11	6.84	8.63
	GF2-GF1	57.15	27.69	42.77
China	GF1-SF	-32.42	-30.59	-30.84
	GF2-SF	-44.30	-41.53	-41.89
	GF2-GF1	-17.58	-15.76	-15.98
Japan	GF1-SF	-3.69	-8.60	-7.58
	GF2-SF	-16.60	-20.53	-18.73
	GF2-GF1	-13.40	-13.05	-12.06
Taiwan	GF1-SF	22.71	21.50	19.64
	GF2-SF	3.21	4.47	4.10
	GF2-GF1	-15.89	-14.01	-12.99

Note: SF: statistical forecasts; GF1: group forecasts in Round 1; GF2: group forecasts in Round 2.

Table 2. Wilcoxon Signed Rank Test Results Evaluated by APE

H_0: test if 0 H_1: test if <0	Test 1 (APE_{GF1}–APE_{SF})	Test 2 (APE_{GF2}–APE_{SF})	Test 3 (APE_{GF2}–APE_{GF1})
Positive ranks (*T*)	17	12	7
Z	-0.463	-1.306	-2.643
Exact *p*. (1-tailed)	**0.328**	**0.099**	**0.004**
Effect size (*r*)	-0.06 (Small effect)	-0.17 (Small effect)	-0.34 (Medium effect)

Note: Within each round of Delphi, experts made forecasts for multiple lead times for every individual source markets. For simplicity, these forecasts were treated as independent.

Table 3. Forecasting Performance Evaluated by APE and Percentage Better by Market

Country/ Region	*Quarter*	APE_{SF}	APE_{GF1}	APE_{GF2}	$PB_{(b-a<0)}$	$PB_{(c-a<0)}$	$PB_{(c-b<0)}$
		(a)	(b)	(c)			
Australia	2011Q2	5.01	1.92	3.62	20.0	20.0	80.0
	2011Q3	1.35	4.84	2.26			
	2011Q4	2.99	6.15	5.30			
	2012Q1	0.16	2.55	2.07			
	2012Q2	1.25	4.70	3.65			
	Mean	2.15	4.03	3.38			
China	2011Q2	28.82	18.47	14.27	100.0	100.0	100.0
	2011Q3	26.88	18.20	14.79			
	2011Q4	20.00	11.10	7.56			
	2012Q1	26.87	17.59	15.39			
	2012Q2	38.30	29.85	26.46			
	Mean	28.18	19.04	15.69			
Japan	2011Q2	19.25	15.28	13.01	40.0	80.0	80.0
	2011Q3	4.42	1.99	0.42			
	2011Q4	1.92	4.42	6.46			
	2012Q1	6.49	7.67	5.74			
	2012Q2	11.46	12.58	10.68			
	Mean	8.71	8.39	7.26			
Taiwan	2011Q2	5.47	4.22	5.32	20.0	40.0	80.0
	2011Q3	0.38	1.34	0.08			
	2011Q4	4.39	6.30	4.77			
	2012Q1	9.30	11.55	9.77			
	2012Q2	4.91	6.59	5.29			
	Mean	4.89	6.00	5.05			
UK	2011Q2	3.35	4.33	3.28	0.0	100.0	100.0
	2011Q3	7.03	7.92	6.96			
	2011Q4	7.16	8.11	7.01			
	2012Q1	6.90	7.56	6.26			
	2012Q2	3.19	3.93	2.70			
	Mean	5.53	6.37	5.24			
USA	2011Q2	1.93	0.42	2.09	80.0	20.0	20.0
	2011Q3	2.00	3.74	1.63			
	2011Q4	2.75	1.16	3.10			
	2012Q1	2.92	0.99	3.15			
	2012Q2	0.76	0.75	1.12			
	Mean	2.07	1.41	2.22			
Grand mean		**8.59**	**7.54**	**6.47**	**43.3**	**60.0**	**76.7**
Std. Deviation		0.10	0.07	0.06			

Note: PB denotes the frequency of smaller APE between any of the two forecasts among SF, GF1, and GF2.

Table 4. Regression Coefficients for Bias (Dependent Variable: PE_t)

Market		Constant	*t*	PE_{t-1}	*t*	Results	Preference on Bias	N	Adjust R^2	*F*-statistic
All	SF	0.001	0.093	0.935	(9.070)**	**Unbiased**	Under	24	0.779	82.262**
(group	R1	-0.002	-0.168	0.905	(6.321)**	**Unbiased**	Over	24	0.629	39.951**
forecasts)	R2	-0.001	-0.083	0.846	(5.069)**	**Unbiased**	Over	24	0.518	25.694**
All	SF	0.001	0.434	0.935	(42.280)**	**Unbiased**	Under	480	0.789	1787.6**
(individual	R1	-0.002	-0.740	0.849	(30.767)**	**Unbiased**	Over	432	0.688	946.583**
forecasts)	R2	-0.001	-0.223	0.784	(23.686)**	**Unbiased**	Over	408	0.58	561.045**
Australia	R1	-0.029	(-6.192)**	0.570	(7.089)**	Biased	Over	72	0.418	50.251**
	R2	-0.029	(-10.586)**	0.253	(3.879)**	Biased	Over	68	0.186	15.044**
China	R1	0.054	(2.830)**	0.845	(8.303)**	Biased	Under	72	0.496	68.946**
	R2	0.079	(3.327)**	0.633	(3.832)**	Biased	Under	68	0.182	14.686**
Japan	R1	-0.017	(-1.449)	0.529	(7.589)**	**Unbiased**	Over	72	0.451	57.593**
	R2	-0.010	(-0.836)	0.496	(6.468)**	**Unbiased**	Over	68	0.388	41.835**
Taiwan	R1	-0.044	(-9.452)**	0.550	(9.167)**	Biased	Over	72	0.546	84.032**
	R2	-0.039	(-10.690)**	0.462	(7.950)**	Biased	Over	68	0.489	63.207**
UK	R1	-0.015	(-2.232)*	0.778	(9.771)**	Biased	Over	72	0.577	95.464**
	R2	-0.065	(-7.371)**	-0.116	(-0.811)	Biased	Over	68	0.01	0.657
USA	R1	-0.004	-0.987	0.710	(8.468)**	**Unbiased**	Over	72	0.506	71.707**
	R2	0.014	(4.056)**	-0.005	-0.041	Biased	Under	68	0	0.002

Note: ** and * indicate significance at the 1% and 5% level, respectively.

Table 5. Binomial Test Results (Bias is measured by the number of (F>A) and (F<A).)

	Category	N	Observed Proportion	*p* (2-tailed)
SF	F<A	14	0.47	0.856
	F>A	16	0.53	
	Total	30	1.00	
GF1	F<A	14	0.47	0.856
	F>A	16	0.53	
	Total	30	1.00	
GF2	F<A	11	0.37	0.200
	F>A	19	0.63	
	Total	30	1.00	