INTEGRATING STATISTICAL AND JUDGMENTAL TOURISM DEMAND FORECASTING APPROACHES:

THE CASE OF HONG KONG

LIN Shanshan

Ph.D

The Hong Kong Polytechnic University

2013

The Hong Kong Polytechnic University

School of Hotel and Tourism Management

INTEGRATING STATISTICAL AND JUDGMENTAL TOURISM DEMAND FORECASTING APPROACHES: THE CASE OF HONG KONG

LIN Shanshan

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

January 2013

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

____________________________________ (Signed)

__________LIN Shanshan________________ (Name of student)

Abstract

In the past 4 decades, quantitative forecasting methods have overwhelmingly dominated tourism demand forecasting studies, while qualitative forecasting research has been correspondingly rare, even though judgmental forecasting has often been exercised in many tourism businesses as a routine and on an informal basis. Given the respective strengths and weaknesses of quantitative and judgmental forecasting methods, it seems sensible to integrate them by combining information from multiple sources. Although combining forecasts has attracted broad attention in the general forecasting literature, only a few studies on this topic have appeared in the tourism forecasting literature. Perhaps due to the absence of contextual information in the forecasting and combination process, the integration of statistical forecasts has led to limited improvement in accuracy. Thus, future studies should focus on integrating judgmental input (including contextual information) into statistical forecasts. To date, there has been little research that has comprehensively examined the effectiveness of integrating judgmental and statistical forecasting methods in the tourism context. Compared to the extensive research on the integration of forecasting techniques in other fields, there is a significant gap in the tourism literature.

This study presents the first attempt to develop a research framework for the integration of econometric and judgmental forecasts based on Hong Kong tourism demand data with a view to providing recommendations and suggestions for decision-makers in both the public and private sectors in Hong Kong. The quarterly forecasts of visitor arrivals in Hong Kong from 6 source markets (i.e. Mainland China, Taiwan, Japan, the USA, the UK, and Australia) up to 2015 were generated

using an econometric model, namely the autoregressive distributed lag model - error correction model (ARDL-ECM). The incorporation of experts' domain knowledge into the statistical forecasts by utilizing the Delphi technique via a forecasting support system (Hong Kong Tourism Demand Forecasting System, HKTDFS) improved forecast accuracy of the integration framework. A qualified panel of experts was formed; the panel members were different stakeholders from the government, accommodation, and tourist attraction sectors, and academics from various institutions.

To establish a holistic analytical framework for integrating statistical forecasts with human judgment, both quantitative and qualitative analyses were applied. The quantitative analysis aimed to examine the forecasting performance of statistical and judgmental forecasts from 3 dimensions: accuracy, bias, and efficiency. The research hypotheses were tested by examining the values of the error measures, conducting correlation and regression analyses, and employing statistical tests. Comparisons were made to examine the difference in accuracy among different Delphi rounds, source markets, expert groups, expertise levels, levels of data variability, forecasting horizons, and sizes and directions of adjustments.

The findings suggested that, on average, statistical forecasts adjusted by the Delphi experts improved forecast accuracy for all of the 6 markets. The results showed that the consensus group forecasts in the final round of the Delphi survey provided significantly more accurate forecasts than those of the initial statistical forecasts and the simple average of individual experts' forecasts in Round 1. Although satisfactory accuracy was achieved, the group forecasts were found to be inefficient and biased for some of the individual markets. It was also found that the

industry experts performed better than the academic experts, indicating the value of incorporating contextual knowledge into statistical forecasts.

In-depth interviews were conducted to provide qualitative input to interpret the quantitative findings from the hypothesis tests, examine the underlying assumptions embodied in the experts' forecasting adjustment process, and collect experts' opinions regarding the use of the forecasting system to aid their judgmental adjustments. The interview findings confirmed that compared to the academic experts, the industry experts preferred to use simpler and easier forecasting methods. The experts reached the consensus that given the relative strengths and weaknesses of judgmental and statistical forecasting methods, it is necessary to integrate these 2 types of forecasts in order to make better tourism demand forecasts. According to the experts interviewed, a variety of reasons were identified as being responsible for the accuracy improvement in this study, such as the provision of multiple information cues (e.g. time-series information and non-time series cues), the use of a Web-based forecasting support system, and the use of the Delphi technique to structure and aggregate experts' judgments. Useful recommendations and suggestions were made by the experts to further improve the HKTDFS and to point to future research directions.

Keywords: Judgmental adjustment; statistical forecasts; integration; Delphi; HKTDFS; Hong Kong

Publications arising from the thesis

1. **Lin, V. S.** (2013). Improving forecasting accuracy by combining statistical and judgmental forecasts in tourism. *Journal of China Tourism Research, 9*(3), 325-352.
2. **Lin, V. S**., & Song, H. (2012). A review of qualitative forecasting in tourism, In C. Cooper (Ed.). Contemporary Tourism Reviews, Oxford: Goodfellow Publishers.
3. Song, H. Gao, 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.
4. Song, H., Gao, Z, Zhang, X., & **Lin, S.** (2012). A Web-based Hong Kong tourism demand forecasting system. *International Journal of Networking and Virtual Organizations, 10* (3/4), 275-291.
5. 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.

Acknowledgements

Though it will not be sufficient to express my gratitude with only a few words to all those people who helped me, I would still like to give my sincere thanks to all of them.

First and foremost, I sincerely thank my chief supervisor, Professor Haiyan Song, for his patient motivation, intellectual guidance, and consistent encouragement in helping me achieve this professional milestone. I much appreciate all his contributions of time, ideas, efforts, and funding to make my PhD experience productive and stimulating. I am also grateful for the excellent example he has provided as a brilliant and successful tourism scholar and professor.

I wish to express my great appreciation to my committee members for their support of this project. I am most grateful to Professor Rob Law, the chair of the Board of Examiners for his valuable comments and insightful questions and to Professor Andrea Saayman (North-West University, South Africa) and Professor Dilek Önkal (Bilkent University, UK), for the critical comments on the draft of the thesis as external examiners. Dr. Kevin Wong and Dr. Jinsoo Lee are also acknowledged for their valuable comments and suggestions on my confirmation report.

My sincere appreciation goes to Professor Paul Goodwin (University of Bath, UK), Professor Eric Girardin (GREQAM, Groupement de Recherche en Economie Quantitative d’Aix-Marseille, France), Dr. Gang Li (University of Surrey, UK), Professor Stephen Witt (University of Surrey, UK), Dr. Chenguang Wu (Sun Yat-sen University, China), Ms Jing Bai Song, Dr. Xinyan Zhang (School of Professional Education and Executive Development, The Hong Kong Polytechnic University), Dr. Zhandong Xu (Dongbei University of Finance and Economics, China), and Dr. Carey Goh, for their encouragement, assistance, and support during the course of my PhD study. My special thanks also go to all of the Delphi panellists for their support and time commitment during the Delphi surveys.

I wish to convey my sincere thanks to the School of Hotel and Tourism Management (SHTM) at The Hong Kong Polytechnic University (PolyU) for offering me an ideal environment in which I felt comfortable and free to concentrate on my study. My time at PolyU was made enjoyable in large part due to the friends and colleges that became a part of my life. The Research Student Group in SHTM has been a source of friendships as well as good advice and collaboration. I also remain thankful to the administrative staffs in SHTM who always provided me with helpful support during my study.

I owe a lot to my two beloved master supervisors, Professor Xuemei Bai and Professor Songshan Zhao (Dongbei University of Finance and Economics, China), who encouraged and helped me at every stage of my personal and academic life, and longed to see this achievement come true.

Lastly, I would like to express my deepest appreciation to my respectful grandparents and parents for their constant love and immeasurable sacrifice, my three dearest sisters, Jingjing for undertaking the editing work, Yunyun and Rongrong for their love and laughs, and most of all for my beloved husband Mr. Huazhen He for his understanding, encouragement, patience, and dedicated love. My family support was a great encouragement that accompanied me throughout this journey. I owe my every achievement to all of them.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Part I:

GF1: Consensus group forecasts in the first round

GF2: Consensus group forecasts in the second round

$GF1_{median}$: Consensus group forecasts by median in the first round

$GF2_{median}$: Consensus group forecasts by median in the second round

$GF1_{mean}$: Consensus group forecasts by mean in the first round

$GF2_{mean}$: Consensus group forecasts by mean in the second round

R1 (Round 1): The first round of the Delphi survey

R2 (Round 2): The second round of the Delphi survey

SF: Statistical forecasts (baseline forecasts)

Part II:

China: China (PRC) or Mainland China

Hong Kong: Hong Kong SAR

Macau: Macau SAR

UK: United Kingdom

USA: United States

Chapter 1 : Introduction

Since the end of Second World War, international tourism has become increasingly accessible to the public. The post-war economic recovery of the industrialized countries led to rising disposable income and free-time availability, both of which are fundamental to engaging in tourism (e.g. in terms of time and money), and stimulated the rapid development of tourism activities. As the foundation of all tourism-related business decision-making processes, the state of tourism demand is the key determinant of business profitability. Undoubtedly, accurate forecasts of tourism demand are essential for establishing effective tourism strategies or plans in the tourism industry, particularly given the perishability of tourism products and services.

Due to the fact that tourism demand is a ubiquitous and growing phenomenon throughout the world today, those public and private organizations that seek to serve and manage this demand need to minimize the risk of future failure, a need that is intensified by the special characteristics of tourism demand and supply. Thus, it is of significant importance, both from the theoretical and the practical perspective, to accumulate knowledge concerning the pattern of tourism demand and its future trends.

The rest of this chapter briefly presents the historical development and contemporary trends of the international tourism industry with a focus on Hong Kong, summarizes the main characteristics of tourism demand, and points out the significance of understanding tourism demand patterns and the necessity of generating accurate tourism forecasts. The discussion further shows the need for incorporating human judgment into statistical forecasting methods in tourism demand forecasting. Research hypotheses are developed to fulfil the research objectives of this study. Finally, the theoretical and practical contributions of the present study are outlined.

1.1 Background

Over the past six decades, tourism has experienced extraordinary growth and diversification to become one of the important economic sectors and the most popular social activity of our time across the world (World Tourism Organization [UNWTO], 2011). International visitor arrivals have enjoyed exponential growth since the 1950s, from 25 million in 1950, to 276 million in 1980, 436 million in 1990, 683 million in 2000, and reached 990 million in 2011, representing an annual increase of 6.3 per cent from 1950 to 2011 (UNWTO, 2012a). Nowadays, tourism contributes directly to 5 per cent of the world's GDP, accounts for one in 12 jobs globally, and is a major export sector in many countries (UNWTO, 2011).

Despite its rapid growth, the tourism industry has suffered from multiple changes and shocks, including man-made crises, natural disasters, and economic crises (UNWTO, 2011). For example, 2011 was a year marked by persistent economic turbulence, major political changes in the Middle East and North Africa, and a devastating earthquake in Japan. Since that turbulent year, global tourism has continued to rebound from the setbacks of the 2008−2009 global financial/economic crisis. UNWTO (2011) has projected that the number of international visitors worldwide will surpass 1 billion by 2012, reach close to 1.4 billion by 2020, and be close to 1.8 billion by 2030. In the period 2010−2030, international visitor arrivals are forecast to increase by 3.3 per cent or 43 million a year on average, compared to an average increase of 3.9 per cent or 28 million a year, in the period 1995−2010

(UNWTO, 2011).

Asia Pacific has been the fastest growing destination region over the past three decades, and this trend will remain in the forthcoming decades as countries within this region progressively develop their tourism products and services in tandem with reduced costs and improved international access to the region. International visitor arrivals to Asia Pacific are predicted to increase by 331 million in two decades, from 204 million in 2010 to 535 million in 2030 (UNWTO, 2011). It is projected that, in relative terms, South Asia will be the fastest growing subregion, with an annual growth of 6 per cent, and Northeast Asia will be the fastest growing subregion in absolute numbers (UNWTO, 2011). By 2030, Northeast Asia will become the most visited subregion, receiving an estimated 293 million visitors. Southeast Asia is expected to receive 187 million visitors thus becoming the fourth most visited subregion, followed successively by Central and East Europe, the Middle East, and North America.

Within the Asia Pacific region, Hong Kong is a famous tourist destination which features a unique fusion of cultures and a great variety of travel experiences and contributes significantly to both regional and global tourism development. In 2011, Hong Kong continued to rank tenth in international tourism receipts (US$27.7 billion) across the world and second within the Asia Pacific region, after Mainland China (UNWTO, 2012b). China is forecast to be one of the leading tourist receiving countries by 2030, and Hong Kong, if treated as a separate entity, will become one of the main destinations (UNWTO, 2011). In 2011, in terms of volume of international visitor arrivals, Hong Kong came third after China and Malaysia in Asia Pacific and second in Northeast Asia. UNWTO (2001) predicted that in 2020, Hong Kong (57 million arrivals) would still be the second most visited destination in

East Asia and the Pacific after China (130 million visitors).

The development of Hong Kong’s tourism industry over the past 30 years has been remarkable (see Figure 1.1): total visitor arrivals in Hong Kong recorded an average increase of 9.8 per cent per annum, from 3.7 million to 41.9 million, over the period 19852011 . The average annual growth rate of international visitor arrivals has gradually slowed down in recent decades, from 12.2 per cent per annum in the pre-handover period to 9.8 per cent per annum after Hong Kong’s return to the Mainland, reflecting the fact that the Hong Kong tourism industry has been greatly affected by the remarkable economic and political changes that occurred after its return to China in 1997. The Asian financial crisis, together with the bird flu epidemic caused a further deterioration in the situation, resulting in a sharp 13.1 per cent fall in visitor arrivals in 1997 and a continuous drop of 9.9 per cent in 1998. Hong Kong’s tourism industry suffered partially because of the pre-handover boom and also because of the fact that the Hong Kong dollar is pegged to the US dollar. The weaker Hong Kong dollar has been a favourable factor in encouraging more visitors from other Asian markets in recent years.

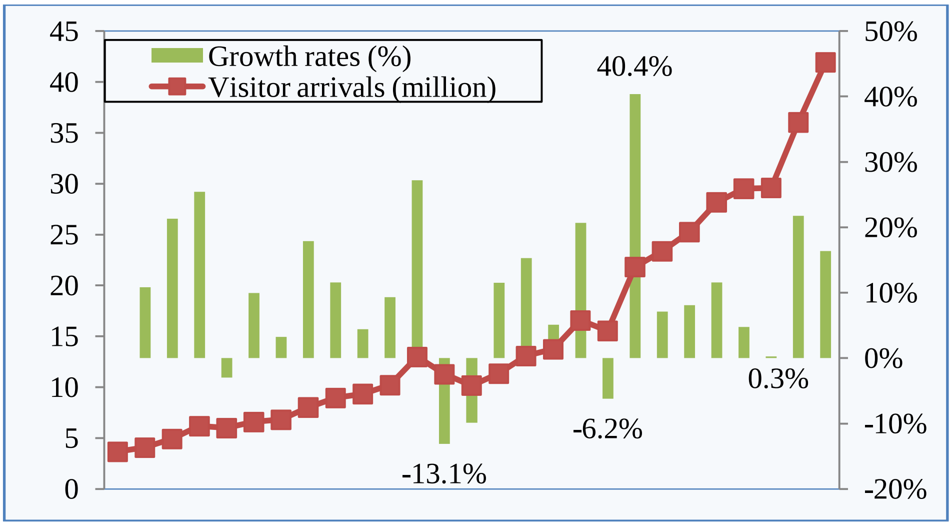


Figure 1.1 Visitor arrivals in Hong Kong 1985–2011

Source: Hong Kong Tourism Board, HKTB (2012a).

Since 1997, Hong Kong's tourism industry has been badly devastated by a series of disasters (e.g. the September 11 terrorist attacks in 2001, the outbreak of the SARS epidemic and bird flu in 2003) which has had a negative impact on the whole industry. Total tourism demand in Hong Kong has been growing since 2003, when the number of visitor arrivals reached 21.8 million, and a rise of 40.4 per cent compared to the previous year was recorded in 2004. This sharp growth was largely driven by a dramatic increase in visitors from Mainland China who were allowed to visit Hong Kong under the Individual Visit Scheme (IVS)[1] in 2003. The opening of the Disneyland theme park in Hong Kong in 2005 also facilitated the boost in tourism, particularly among young Chinese tourists and their families, in spite of the negative press reports about the new development.

Being one of the world's most open economies, Hong Kong is especially vulnerable to financial turmoil in the world economy. Boosted by Hong Kong's co-hosting of the Beijing Olympics Games and the Paralympic Games, the Hong Kong tourism industry chalked up a strong year-on-year growth of 8.9 per cent in visitor arrivals for the first half year of 2008. However, the surge started to stagnate in September 2008 following the onset of the global financial crisis. Nevertheless, Hong Kong received 29.5 million visitors by the year's end, a modest 4.7 per cent rise over 2007. Given the strong momentum from Mainland China, the Hong Kong tourism industry was overshadowed but less affected by the global economic meltdown and the outbreak of human swine influenza (H1N1). Visitor arrivals to Hong Kong registered a minor growth of 0.3 per cent in 2009 (HKTB, 2012b) while

[1] The Individual Visit Scheme (IVS) was first introduced in four Guangdong cities (Dongguan, Zhongshan, Jiangmen, Foshan) on 28 July 2003 as a liberalisation measure under the Closer Economic Partnership Arrangement (CEPA) (The Government of the Hong Kong Special Administrative Region, 2008). The scheme allows residents of these cities to visit Hong Kong in their individual capacity with 7-day visas issued by the Mainland's Public Security Bureau. The coverage of the Scheme has now been expanded to 49 Mainland cities.

total tourism expenditure associated with inbound tourism achieved a slight growth of 3.2 per cent (HKTB, 2012b).

Sustaining its recovery and upward trend in 2010, Hong Kong's tourism industry continued to achieve significant growth in 2011, led by the rise in travel aspirations resulting from the continued improvement of the global economy and the strengthening of most currencies against the Hong Kong dollar (The Legislative Council Commission, 2012). In 2011, total visitor arrivals hit a new record of 41.92 million (The Government of the Hong Kong Special Administrative Region, 2012), representing a year-on-year increase of 16.4 per cent. The Mainland market remained the key growth impetus, with visitor arrivals from this market surging by 23.9 per cent to 28.10 million, accounting for 67.0 per cent of the total.

However, the growth momentum in terms of visitors from other source markets, especially those from the USA and Europe, was far less impressive. Visitor arrivals from long-haul[2] markets only edged up by 1.7 per cent to 4.77 million while there was a moderate increase of 4.6 per cent to 9.05 million in visitor arrivals from the short-haul markets (The Legislative Council Commission, 2012). In tandem with the surge in visitors, visitor spending as reflected by exports of travel services had another year of strong growth in 2011, thereby providing an important growth driver for the local economy at a time when external trade was faltering as a result of the fragilities in the advanced economies (The Government of the Hong Kong Special Administrative Region, 2012). In 2011, t*otal expenditure associated with inbound tourism* soared by 20.5 per cent year-on-year to HK$253 billion (HKTB, 2012a). The per capita spending of both overnight and same-day visitors also surged by 9.0 per cent to HK$7,333 and by 4.0 per cent to HK$1,920, respectively (The

[2] Short-haul markets include Taiwan, Japan, South Korea, Singapore, Malaysia, Thailand, Indonesia, and the Philippines, while long-haul markets include the USA, Canada, the UK, Germany, France, and Australia (The Legislative Council Commission, 2012).

Legislative Council Commission, 2012).

Tourism, being one of the four traditional pillar industries (the other three being financial services, trading and logistics, and producer and professional services), is not only the key driving force of Hong Kong's economic growth but also provides critical impetus to investment and employment in the local economy. The impact of travel and tourism runs deep into the economy. It is not just about the money that visitors spend on travel, accommodation, activities, and souvenirs; but by its very nature, the industry stimulates the engagement and collaboration of communities, tourists, governments, local suppliers, and businesses throughout the supply chain (WTTC, 2011). The total contribution of the travel and tourism industry to GDP was amounted to US$4.95 billion, or 8.3 per cent of total GDP, in 1988 and surged to US$36.96 billion, or 12.6 per cent of total GDP, in 2011. The estimated 0.46 million people in Hong Kong whose jobs are supported by the travel and tourism industry (0.23 million of whom work directly in the industry) all spend a proportion of their own income on goods and services from all parts of the economy (WTTC, 2012). Furthermore, demand for travel and tourism stimulates investment. In 2011, 9 per cent of total capital investment, or some US$4.79 billion, was driven by the travel and tourism industry.

When examining the top 10 tourism-generating countries and regions of Hong Kong during the period 2005−2011, China, Taiwan, the USA, Japan, Macau, South Korea, the UK, Australia, the Philippines, and Singapore accounted for nearly 90 per cent of total arrivals (see Figure 1.2). With the exception of the UK, the USA, and Australia, the main source markets for Hong Kong tourism are all in Asia, with Mainland China being the predominant market. Mainland Chinese make up by far the largest segment of visitors, typically representing more than half of the visitors entering Hong Kong (averaging 58.6% from 2005−2011), according to HKTB (2011). The continued expansion of the Mainland economy and the appreciation of the RMB versus the Hong Kong dollar attracted many Mainland visitors to pay consumption visits to Hong Kong in 2011, causing arrivals from this source market to surge by 23.9 per cent year-on-year to 28.10 million. Of these, 65.3 per cent or 18.34 million, travelled via the IVS, representing a year-on-year increase of 28.8 per cent (The Government of the Hong Kong Special Administrative Region, 2012). More than 30 per cent of these visitors were Shenzhen permanent residents travelling on the one-year multiple-entry endorsement under the IVS. The number of business visitors from the Mainland, many of which were meetings, incentives, conventions and exhibitions (MICE) arrivals, also increased steadily.

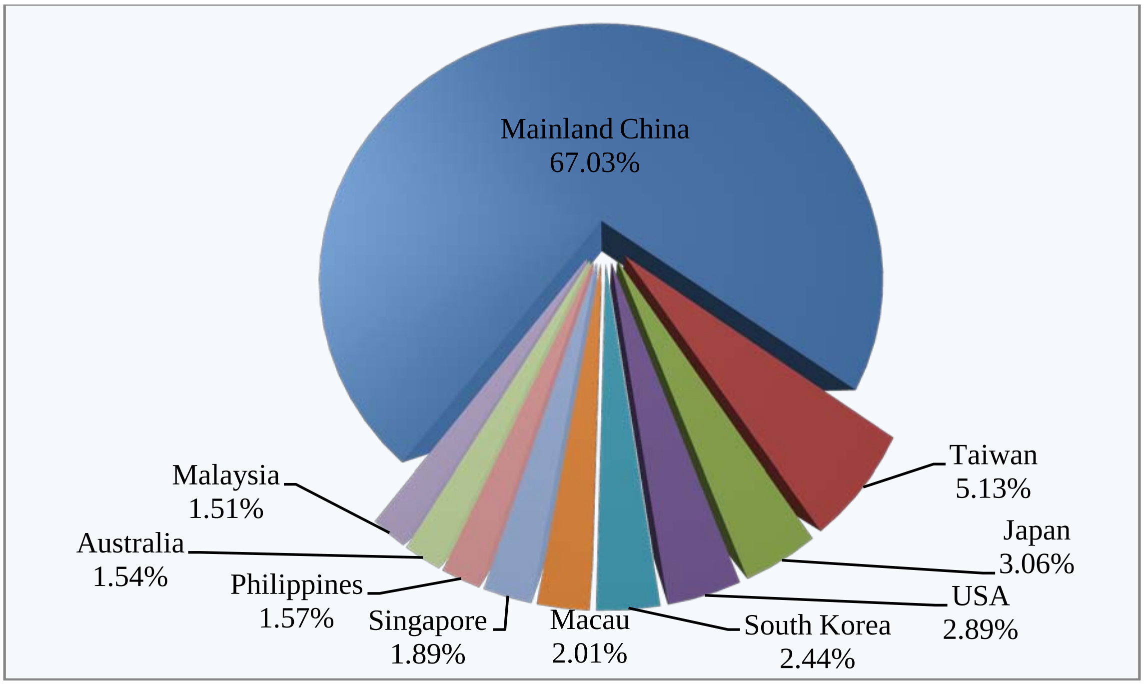


Figure 1.2 Top 10 tourism source markets of Hong Kong in 2011

Source: HKTB (2011).

Hong Kong is beginning to lose its popularity with Taiwanese visitors despite the ease and convenience of travelling between Taiwan and Hong Kong. In 2011, the total number of arrivals from this market was maintained at a similar level to (2.15 million). Same-day business arrivals from Taiwan continued to decrease with the further expansion of direct cross-strait flights. HKTB (2012b) reported a total of 1.29 million arrivals from Japan in 2011, just 2.4 per cent less than the number in 2010. The earthquake and subsequent nuclear plant crisis in Japan in March 2011 dampened outbound travel sentiment in Japan, and this led to a huge drop in arrivals to Hong Kong in the second and third quarters of the year. Positive growth was registered in the fourth quarter as a result of the impact of the earthquake beginning to wear off, coupled with the strong Japanese yen.

1.2 Problem Statement

1.2.1 The need for tourism demand forecasting

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 6 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). Virtually all policy analysis and planning problems require forecasts of future demand. Sound tourism demand forecasts can be helpful to marketers and managers in reducing the risk of decisions made with respect to the future; for example, tourism marketers can use demand forecasts to set marketing goals, explore potential markets, and simulate the impact of future events on demand. More specifically, the demand predictions may include the travel volumes, the market share of various destinations, the hotel room nights, or the number of passengers flying between two destinations.

The rapid expansion of international tourism has motivated a growing interest in tourism demand studies. The last 4 decades have witnessed significant developments in tourism demand analysis with respect to the depth of theoretical foundations, the diversity of research interests, and advancements in research methodologies. Tourism demand studies mainly focus on two aspects: the analysis of the effects of various determinants and the provision of accurate forecasts of the future tourism demand. This study focuses on the latter.

Tourism demand forecasting is a prerequisite to the decision-making process in many organizations in the private and public sectors because it is useful for improving the efficiency of the decision-making process. Managers can understand the changes taking place in the economy better by undertaking economic forecasting. Put simply, accurate tourism demand forecasts can improve the efficiency of business, increase profits, and strengthen economies (UNWTO and ETC, 2011). Forecasting tourism volume is particularly important because tourism volume is an indicator of demand that provides basic information for subsequent planning and policymaking (Chu, 2008). Arguably, all industries are interested in forecasting tourism volume since this helps to improve the allocation of scarce resources to avoid shortages or surpluses (Burger, Dohnal, Kathrada, & Law, 2001). Any information concerning the future evolution of tourism flow is of great importance to hoteliers, tour operators, and other industries related to tourism or transportation as such information allows them to adjust their policies and corporate finance. For instance, having a good idea of the number of tourists visiting a particular country, region, town, attraction, or hotel in a given time period helps tourism managers to plan much more effectively. If demand is predicted to increase, more staff can be hired, more excursions arranged, accommodation capacity increased, and so forth. Furthermore, the increasing impact of tourism receipts on the national balance of payments and economic growth makes forecasts of inbound flows from each source market more essential to the government for planning and marketing purposes. This makes forecasting a highly important field for the tourism industry, which needs accurate demand forecasts to plan effectively from season to season, year to year (UNWTO and ETC, 2011). Therefore, visitor arrival variables have been the most frequently researched measure of tourism demand over the past few decades (Song & Li, 2008).

Due to the key role of demand as a determinant of business profitability, projections of future demand form a very important basis in all business planning activities. Nevertheless, this risk reduction is more acute in tourism industry for the following reasons (Frechtling, 2001; Song & Guo, 2008; Tsamakos, Giaglis, & Kourouthanassis, 2002).

(1) Tourism products are highly perishable. Tourism products are time critical, have to be consumed in a specific space and period, and cannot be stored in any way. For example, an unsold hotel room, an unused aircraft seat, or a vacant concert seat can render no value to the suppliers after the specific offer time has elapsed. This implies high risks for the suppliers and also puts a premium on shaping demand in the short run and anticipating it in the long run.

(2) People are inseparable from the production-consumption process. In a broad sense, the production of a tourism product/service appears at the same time as its consumption. In addition, much of this process involves interactions between suppliers and consumers, such as travel agency staff and tourists. Tourism products are also regarded as good experiences as one of their main components is tourist

satisfaction. A tourist acquires experiences while interacting with a new environment, and his/her experiences help to attract and motivate potential customers.

(3) Customer satisfaction depends on complementary services. The tourism product usually consists of a set of elementary products which, unlike a manufactured product, cannot be provided by a single enterprise; for instance, an airline supplies seats, a hotel provides rooms and restaurants, travel agents make bookings for the stay and sightseeing, and so forth. Forecasting can help to ensure that these complementary services are available to satisfy the needs of future visitors.

(4) Leisure tourism demand is extremely sensitive to natural and human-made disasters. Most holiday and vacation travellers are motivated by the desire to seek refuge to get rid of the pressures from their routine environment and people. Nowadays, numerous alternatives are available for spending leisure time in ways other than travelling. In fact, tourism is perhaps more vulnerable than any other industry to demand fluctuations stemming from seasonal, economic, political, social, and other such factors. The reasons for these fluctuations can be attributed to natural disasters (e.g. earthquake, typhoon, flood, disease) and human-made crises (e.g. war, terrorist attacks, crime, strikes). The ability to predict the possible impacts of such events on tourism demand can help to minimize the adverse effects on tourism-related businesses.

(5) Tourism supply requires large, long lead-time investment in plant, equipment, and infrastructure. Tourism investment, particularly in destination infrastructures, requires long-term financial commitments; for example, it may take years for a new airport to be formally launched. The opportunity costs could be very high if an investment project fails to fulfil its expected capacities. Therefore,

forecasts of the long-term demand for tourism-related infrastructure should be produced to avoid such losses.

1.2.2 Characteristics of tourism demand forecasting

Yeoman (2008) claimed that forecasting is essentially a hazardous exercise and that projecting future international tourism flows has become more difficult over time. Special events, such as natural disasters, wars and diseases, cause only temporary interruptions; however, other factors, such as the nature of a political regime and restrictive legislation, are likely to yield more permanent changes. The nature of tourism demand presents a number of special challenges to forecasters and practitioners that do not afflict their counterparts in other industries (Frechtling, 2001).

(1) *Seasonality*

Seasonality has long been recognized as one of the most prominent and worrisome facets of the tourism industry, and it may be the most typical characteristic of tourism on a global basis. Seasonality generally indicates the phenomenon of fluctuations in demand or supply in the tourism industry due to factors such as the climate, institutional patterns (e.g. school or calendar holidays), lifestyles, special events, and the like (Allcock, 1989; Chung, 2009; Nadal, Font, & Rossello, 2004). It can be defined as a cyclical pattern that more or less repeats itself each year (Jang, 2004). Alternatively, it can be seen as a “temporal imbalance in the phenomenon of tourism, which may be expressed in terms of dimensions of such elements as the number of visitors, expenditure of visitors, traffic on highways and other forms of transportation, employment and admissions to attractions” (Bulter, 2001, p. 5).

One of the most exhaustive reviews of seasonality conducted so far is that offered by BarOn (1975), who examined this issue in relation to 16 different countries using data covering a period of 17 years. Since then, many scholars have followed BarOn's lead and have continued to further investigate this issue by either briefly extending the discussion in the book or initiating other topics as their main focus (Bulter, 2001). The overwhelming consensus from such writings is that "seasonality is a problem" and "is something to be overcome, or at least modified and reduced in effect" (Bulter, 2001, p. 10). Yacoumis (1980) described seasonality as "an almost universal problem, varying only in the degrees of its acuteness from one country to another" (p. 84).

It is generally agreed that the fluctuations of seasonality are attributed to two main factors: natural and institutional (BarOn, 1975). The former is usually caused by regular climatic changes throughout the year (e.g. temperature, rainfall, snowfall, and sunlight), and the latter consists of factors that reflect the social norms and practices of society (e.g. religious, school, and public holidays) (Butler, 1994). Three more causes of seasonality, namely social pressure or fashion (taking the waters at spas or hunting on country estates among the privileged elites), sporting season (snow skiing or surfing), and inertia or tradition, were added by Butler (1994).

Many studies agree that seasonality may result in severe economic and social issues, such as an unstable labour market caused by temporal employment in a destination (BarOn, 1975; Chung, 2009; Nadal, Font, & Rossello, 2004; Yacoumis, 1980). To understand seasonality better, Lundtorp (2001) suggested examining the fluctuations in tourism reflected in basic measures, such as number of visitors, not only on an annual basis but also by month, week, and day. Expenditure levels are also an important measure of seasonal demand (Nadal, Font, & Rossello, 2004).

(2) External interventions

While a variety of factors can influence tourism demand, one of the most perceptible contributions comes from external shocks, such as changes in government regulations, economic conditions, political environment, and health and safety conditions in either the tourism origin countries or the destinations; these sorts of shocks partially contribute to the long-term volatility of tourism demand.

As for the tourism industry in Hong Kong, a series of events has occurred in the past few decades. The worldwide petroleum shortages in 1973−1974 and 1979−1980 temporarily reduced international and domestic tourism demand in many countries. After its return to China in 1997, Hong Kong’s tourism industry experienced a tough period due to the political and social economic changes occurring in its society. The financial crisis in 1997 depressed international tourism demand in the Asia Pacific region. The outbreak of SARS in 2003 seriously affected the tourism industry throughout the world, particularly for destinations in Asia. More recently, the global economic downturn that started in 2008 caused substantial losses in international visitor arrivals and tourism receipts. This negative trend intensified during 2009 and was exacerbated by the H1N1 influenza virus.

Notwithstanding the negative impacts brought by the above events, other external interventions such as mega events, can cause a noticeable growth in tourism demand. Roche (2000) defined mega events as “large-scale cultural (including commercial and sporting) events which have a dramatic character, mass popular appeal and international significance” (p. 1). Such events as the Olympic Games, the

FIFA World Cup, world fairs, and other international sport championships are not only likely to attract an increasing global audience (Fourie & Santana-Gallego, 2010; Getz, 2008; Horne & Manzenreiter, 2004) but can also shape world tourism patterns, highlight new tourism destinations, and create "lasting legacies" in the host cities or countries.

As such events and disasters are virtually impossible to forecast, their possible impacts on tourism are even more obscure. It is therefore vital to incorporate the impacts of such external interventions into a forecasting model in order to improve the robustness and fitness of the model and hence produce more accurate forecasts.

(3) Complexity of tourism behaviour

A wide range of reasons motivate people to go travelling, the main three being leisure tourism (e.g. holidays, health and fitness, sport, education, culture and religion, and social and spiritual purposes), visiting friends/relatives (VFR), and business tourism (Rowe, Smith, & Borein, 2002). Business travel in the tourism industry embraces a wide variety of forms, including corporate meetings, conferences and conventions, exhibitions and trade fairs, training courses, new product launches, press conference presentations, and lobbying government officials (Rowe, Smith, & Borein, 2002; Swarbrooke & Horner, 2002). Furthermore, business people travelling for business purposes on one day can plan for their individual trips on the next, but the motivations for business travel differ significantly from the motivations for leisure travel.

The purposes of leisure travel are different from those of business trips. If all trips with different purposes were combined into a single tourism demand series, it would be difficult to obtain the best forecasting models. Turner, Kulendran, and

Pergat's (1995) study showed that more accurate forecasts can be obtained by distinguishing between the following series: holiday, VFR, business, and other purposes. However, disaggregated data may not always be available (Frechtling, 2001).

The travel and tourism industry is divided into six key areas: travel agents, transportation, attractions, tour operators, tourist information and guiding services, accommodation, and catering (Rowe, Smith, & Borein, 2002). These components do not work separately but rather in an interactive manner. Forecasters do not have a clear concept of the "travel product" that a particular tourist/visitor seeks and how the aforementioned components affect his or her purchasing decisions with regard to complements and substitutes. Moreover, no consensus has been reached on establishing a sound theoretical foundation for tourism demand. Managers and forecasters are not even certain about what drives and determines families' purchases of vacation travel packages.

(4) Measures of tourism demand

Tourism demand is regarded as "a measure of visitors' use of a good or service" (Frechtling, 2001, p. 4). There is a wide choice of forecast variables measured in various units, such as national currency, arrivals, nights, days, distance travelled, and passenger seats occupied. Frechtling (2001, p. 15) categorized the alternative measures of tourism activity into six groups: (a) visitors, measured in "number of people travelling away from home"; (b) visitor expenditure, measured in "total money spent purchasing goods and services related to the trip"; (c) visitor-nights, measured in "total nights visitors spend away from home"; (d) visitor-miles/kilometres, measured in "distance travelled while away from home"; (e) visitor parties, measured in “groups of people travelling away from home together”; and (f) market size, measured in “number of people travelling away from home once or more in a year”.

In terms of statistical availability and consistency between data sources, the first two categories, tourist arrivals and tourist expenditure, together with their derivatives, such as (a) tourist participation rate derived from tourist arrivals divided by population of the origin country/region, and (b) tourist expenditure per capita derived from total tourist expenditure divided by population, have been the most frequently used tourism demand measures in empirical studies over the past 4 decades (Song, Li, Witt, & Fei, 2010; Song, Witt, & Li, 2009; Witt & Witt, 1995). Based on a comparative study of three reviews from Crouch (1994), Lim (1997), and Li, Song, and Witt (2005). Song et al. (2010) summarized the applications of various tourism demand measures and found that tourist arrivals and tourist expenditure with their derivatives had dominated tourism demand modelling and forecasting studies. A more thorough discussion can be found in Calantone, Benedetto, and Bojanic (1987), Lim (1999, 2006), and Song, Witt, and Li (2009).

(5) Data collection

Generally, the longer the time series available, the more likely it is that forecasting models will capture the patterns of the tourism activity to be forecast. The majority of forecasting methods require a minimum sample size of 5 years or more for annual data. For instance, five data points are required to generate a 1-year-ahead forecast (Frechtling, 2001), while, according to this rule, at least 10 observations are required to generate a 2-year-ahead forecast. For monthly data, if people want to forecast 1 year ahead, this requires at least five complete years of a monthly series. However, few cities or destinations, particularly newly developed destinations, have such records. This rule should therefore be regarded as a minimum requirement as some sophisticated quantitative forecasting methods need more data points to provide reliable estimates.

It is ideal that the time series over the period we wish to study is adequate and complete. In many cases, the historical data series are either discontinuous or have anomalous values (e.g. outliers). For example, the Hong Kong Tourist Association maintained a time series of tourism receipts in Hong Kong from international tourists until 2001; however, a new methodology for compiling and presenting tourism expenditure statistics has been adopted since 2002, producing a new time series from 2002 onwards. Apparently, there are two different time series here: 1985−2001 and 1998−2009. Yet, the data were published as a single time series showing a remarkable drop in tourism expenditure in 1998. Using this series for modelling and forecasting would show this drop to be the result of changes in tourist behaviour rather than a change in estimation methodology. To solve this problem, one solution is to deal with the consistent time series of 1998 to the present, while the other solution is to try to adjust the old series to make it comparable to the new series. As suggested by Frechtling (2001), people should not just simply shift the old series up or down by an amount to splice it in with the new one as this splicing assumes that the old method captures tourist behaviour as well as the new one; however, there is no evidence to support such an assumption.

Another problem in the historical data is the existence of extreme values and outliers which usually deviate far from the established pattern of the time series to be forecast. The value for 2003 (i.e. denoting the impact of SARS) in Figure 1.1 is an outlier in the visitor arrivals series. It is easy to identify such outliers by examining past experience of similar data or visualizing the graphs (e.g. frequency histogram normal probability plots) or, more formally, by conducting a normality test. Most statistics experts argue against adjusting outliers in some way (e.g. by using dummies to detect the sudden change). Alternatively, the measurement error could be a solution if there are no appropriate explanations.

1.2.3 Methods and models of tourism demand forecasting

It is worth noting the difference between *a forecasting method* and *a forecasting model*. As clarified by Frechtling (2001), a forecasting *method* is "simply a systematic way of organizing information from the past to infer the occurrence of an event in the future" while a forecasting *model* "is one expression of a forecasting method" (p. 21). More specifically, a forecasting model may be single equation or a group of related equations. The existing literature suggests that there are two categories, qualitative and quantitative forecasting methods, in tourism studies. This division is largely based on the availability of historical time-series data.

Qualitative methods (also called *judgmental methods,* which is the term adopted in this study) refer to a variety of nonscientific techniques including intuition, used to project future developments. Past information on the forecast variable of interest is organized by using experts' judgments or opinions rather than mathematical rules. One of the best-known judgmental forecasting methods is the Delphi technique, which involves the formal and structured soliciting of judgments concerning a given forecasting problem from a group of knowledgeable experts.

Quantitative methods quantify past information about a phenomenon by applying mathematical rules which take advantage of the underlying patterns and relationships in the data. These methods assume that some elements of past patterns will continue into the future to some extent (Makridakis, Wheelwright, & Hyndman, 1998). There are two major subcategories of these methods: time series (or extrapolative, noncausal) methods and econometric (or causal) methods.

The first subcategory examines trends and cycles in the historical data of a particular variable and then uses mathematical techniques to extrapolate them into the future. These methods assume that the observed trend of the variable will continue for some reasonable period into the future. This is often a valid assumption when forecasting short-term horizons, but it falls short when making medium- and long-term forecasts.

The second subcategory attempts to identify a cause-and-effect relationship between a measure of tourism demand and its influencing factors (e.g. price, income), support policy evaluation, and strategy making and to predict future trends in tourism development. The principal objective of this method is to discover the major determinants (or the explanatory, independent variables) that can affect the forecast variables (or the dependent variables) and to select an appropriate mathematical function form to portray this relationship.

A third branch in forecasting methodologies, artificial intelligence (AI) technology, is a newly emerging and rapidly developing area in the tourism forecasting literature. These methodologies can be used as either causal or noncausal forecasting methods depending on whether any influencing factors of tourism demand are considered. The frequently used AI forecasting techniques include neural networks, rough sets, genetic algorithm, support vector regression, fuzzy logic, grey theory, and their combinations (Li, 2009). The advantage of applying AI forecasting techniques is that they are relatively less restricted by data property requirements (e.g. stationary and normal distribution), and this enables them to provide more accurate forecasting results compared to their traditional counterparts (Li, 2009). However, the lack of theoretical foundation in social sciences such as economics leads to difficulty in interpreting tourism demand from the economic perspective and therefore provides very little support in policy evaluation (Song & Li, 2008).

The challenges of producing successful forecasts are far more than just the technical difficulties of developing an accurate model. The selection of the most appropriate model involves consideration of the following four factors identified by Stynes (1983, cited in Smith, 1995): (a) the organizational environment, (b) the decision-making situation, (c) the existing knowledge, and (d) the nature of the phenomenon being studied.

A review of the requirements and characteristics of the most frequently used forecasting methods in the general forecasting literature is tabulated in Table 1.1; it reveals the fact that no single model is best on all criteria. Two important criteria, accuracy and precision, are not included in this table because it is hard to decide which method is always superior to others under all circumstances.

In general, quantitative methods exhibit better performance in predicting future revenues than judgmental methods (Armstrong, 2001e) when large changes are not expected or the historical data are adequate; however, if little or no data exist, the use of judgmental methods is suggested. In addition, judgment is incorporated as part of the forecasting process even in quantitative applications; for instance, providing inputs or deciding which quantitative method to use involves human judgments.

It is clear that no one forecasting technique is superior for all occasions. The choice of technique often requires trade-offs between convenience ("what's easy"), market popularity ("what others do"), accuracy, precision, time constraints, financial support, and other resources (Smith, 1995). All models are capable of generating good-quality forecasts if they at least satisfy the following three conditions (Smith, 1995): (a) models were appropriately developed and applied, (b) sufficient historical data, and (c) the problem being investigated closely conformed to the implicit assumptions of the selected models.

Armstrong (2001e, p. 376) developed a flow chart (Exhibit 4) to illustrate how to select appropriate forecasting methods. Generally, quantitative methods are more accurate than qualitative methods when enough historical data are provided; causal methods are more accurate than Naive methods when big changes are expected; and simple methods are better than complex methods as they are easier to understand, much cheaper, and seldom less accurate. Before choosing a judgmental approach, forecasters should at least consider the following factors: whether there are large changes, frequent forecasts, and conflicts among decision-makers. On the other hand, when choosing a quantitative method, they should examine the type of data (cross-sectional data or time-series data), the amount of change involved (large or small change), and the prior knowledge about the future relationships.

All of the forecasting methods or models discussed involve a certain degree of combined practice. Substantial evidence supports the argument that combining individual forecasts produces gains in forecast accuracy (Armstrong, 2001a; Bunn, 1988; Fritz, Brandon, & Xander, 1984; Hallman & Kamstra, 1989; Pollack-Johnson, 1995). Combining forecasts provides us with a way to compensate for deficiencies in a forecasting technique. By selecting complementary methods, the shortcomings of one technique can be offset by the advantages of another. There is also a considerable amount of evidence which suggests that adding quantitative forecasts to qualitative forecasts either increases or reduces accuracy (Bunn & Wright, 1991; Flores & White, 1989; Goodwin & Wright, 1993; Lawrence, Edmundson, & O'Connor, 1985; Pereira, Coqueiro, & Perrota, 1989). However, no research has yet disclosed the conditions or methods required for the optimal combinations of forecasts.

Table 1.1 Summary of types and characteristics of forecasting methods

Method	Technical expertise required	Data type/source	Required data precision	Forecasting horizon	Time required for forecast	Type of problem best suited for	Computing resources required
I. Quantitative methods							
1. Time series methods (1) Naive models (2) Moving average (MA) models (3) Autoregressive (AR) models (4) Exponential smoothing models (5) Classical decomposition (X11) (6) Box-Jenkins models ((S)ARIMA)	Low to medium	Time series data	Medium to high	Short	Short	Simple, stable, or cyclic	Minor to moderate
2. Econometric methods (1) Regression analysis (2) Structural econometric models (3) Spatial models	Medium to high	Time series data, plus causal relationships, and change processes	High	Short to medium	Short to medium	Moderately complex with several variables and known, stable relationship	High
3. Artificial intelligent forecasting methods (1) Artificial neural networks (ANN) (2) Fuzzy logic systems (FL) (3) Genetic algorithms (GA) (4) Grey models (5) Expert systems (6) Rough set approach	High	Time series data or cross-sectional data	Medium to high	Short to medium	Short to medium	Moderately complex with several variables	High
II. Qualitative methods							
1. Intentions 2. Opinions 3. Delphi method 4. Traditional meeting 5. Structured meeting 6. Group depth interview 7. Role playing	Low to medium	Expert and experiential data	Low	Long	Medium to long	Complex with known but qualitative relationship and elements of uncertainty	Low

Sources: Frechtling (2001), Hu (2002), and Smith (1995).

1.2.4 The need for integration

In short, the demand for products and services in the tourism industry can be affected by an incredible number of different factors, such as economic conditions and lifestyles, fuel prices, tourist infrastructure, hotel prices, environmental changes, and natural disasters (Kollwitz, 2011). Because of all of these factors, tourism demand in all of its different forms is one of the most difficult variables to forecast (UNWTO and ETC, 2011). There are numerous ways to predict tourism demand, ranging from asking experts to give their gut feelings to highly complicated forecasting models to provide more accurate forecasts.

In an increasingly competitive industry, tourism decision-makers are faced with the necessity of making projections of future demand in the short term despite the limitations of scarcity, volatility, and uncertainty. It is realistic to deduce that in their regular decision-making process, these experts make use of confidential information and qualitative stimuli to estimate future values (Croce & Wöber, 2011); for instance, a hotel manager who has worked for years in the same hotel is able to roughly predict future room demand for a particular period of the year based on his or her prior knowledge of the same period in past years (e.g. room demand during the Christmas season). However, such direct judgments have a number of serious disadvantages compared to statistical methods. It has been proposed that people have limited information processing time and cognitive capacity and thus often use simplifying mental strategies, or heuristics, to cope with the complexities of a judgmental task (Gilovich, Griffin, & Kahneman, 2002; Tversky & Kahneman, 1974). The use of heuristics, however, often involves compromising accuracy for efficiency and speed (Gilovich, Griffin, & Kahneman, 2002).

A consensus has been reached among many researchers and forecasters that judgment should play an important role in forecasting practice (Ghalia & Wang, 2000; Goodwin, 2002; Sanders & Ritzman, 1992; Wright, Lawrence, & Collopy, 1996), but this approach could also be subject to biases and inconsistencies resulting from "cognitive limitations, political influences or confusion between forecasts, targets and decisions" (Parackal, Goodwin, & O'Connor, 2007, p. 343). Forecasters are possibly prone to bias either in the generation of forecasts or the evaluation of outcome or both. Some researchers believe that decision-makers have difficulty in consistently applying their information-processing strategies; for example, Beach and Christensen-Szalanski (1987) contended that issues regarding the quality of human judgment have not been settled and that the commonly held belief that human judgment is poor is not based on convincing data. Thus, as suggested by Butler, Kavesh, and Platt (1974), statistical methods could be used as tools to offer a first approximation of forecasts by utilizing the historical data.

Whereas the judgmental forecasting methods are prone to suffer from a number of biases (e.g. optimism, wishful thinking, lack of consistency, and political manipulation), formal quantitative methods struggle when past data are scarce and also suffer major difficulties in handling special events or significant changes in the environment, such as promotion campaigns or new government policies (Goodwin, 2002). Although complex statistical methods enable managers/forecasters to make optimal use of vast quantities of data and to handle these data consistently, they may also lack transparency, and hence credibility, and deny managers/forecasters the sense of owning the forecasts.

In the past decades, many scholars have extensively reviewed the research into judgmental forecasting (e.g. Armstrong, 2001e; Goodwin & Wright, 1993;

Lawrence, Goodwin, O'Connor, & Önkal, 2006; Parackal, Goodwin, & O'Connor, 2007; Webby & O'Connor, 1996) and have shown a general acceptance of the strategy of adopting combined methods under different conditions instead of favouring a single statistical or judgmental approach. When the historical data are insufficient or large changes are expected to occur in the future, judgmental forecasting methods are preferable; otherwise, statistical forecasting methods are favoured. Virtually all of the statistical techniques used for forecasting require a series of historical data that can be used in preparing the forecasts. Klein and Newman (1980) illustrated that a turbulent environment characterized by discontinuous change may be unpredictable by statistical forecasting techniques that count on the continuity of historical data. Judgmental forecasts are also adopted when there is insufficient time to produce statistical forecasts or the situations are changing so rapidly that statistical forecasts would become useless (Wright & Ayton, 1987). In the existing tourism research, the issue of missing data has been exclusively addressed when using qualitative methods to forecast tourism demand (e.g. forecasting the room demand for a new hotel or the visitor volume of the Shanghai Disneyland resort). In such situations, the managers can devise forecasts based on subjective judgments rather than historical data because such data are unavailable or irrelevant.

However, the contribution of academic research to understanding and assessing the contributions of experts' judgments to tourism forecast accuracy is still considerably limited in comparison with the published studies in the general forecasting literature. Most tourism forecasting research has been devoted to the area of quantitative models (Song & Li, 2008; Witt & Witt, 1995), and it is surprising that the considerable advances in judgmental forecasting achieved in other disciplines have still not received much attention in the tourism forecasting literature. Given the knowledge capital possessed by tourism analysts, the industry could benefit from just one attempt to exploit this resource to achieve more accurate forecasts.

The complexity of tourism behaviour and the volatility of demand and its elasticity to events demand the formulation of models with a high degree of flexibility which challenges the performance of quantitative models. Even in the short horizon, where the impact of uncertainty is limited, demand sensitivity to changes both in the destination and in the origin adds complexity to the modelling exercise. As suggested by the general forecasting literature, such an environment presents sufficient threats to encourage the use of judgmental forecasting methods (Armstrong, 2006). To date, no published study has comprehensively examined the effectiveness of forecasting integration in the tourism context. Compared to the extensive research on the integration of forecasting techniques in other fields, there is a significant gap in the tourism literature. Considering the performance of current tourism forecasting techniques and the importance of tourism demand forecasting for a typical tourist destination like Hong Kong, this gap needs to be seriously addressed. Based on the vast empirical evidence in the literature, it is expected that a pool of forecasts from judgmental and statistical methods will improve accuracy in the context of tourism demand forecasting.

1.3 Research Objectives

The main research objectives of this study are as follows:

1. To develop a research framework for the integration of statistical forecasts and judgmental forecasts based on Hong Kong tourism demand data.
2. To implement the forecasting integration framework developed (objective 1) in forecasting the demand for Hong Kong tourism.
3. To examine the effectiveness of the integration of judgmental and statistical forecasting methods.
4. To propose a research agenda for further developing a more effective and supportive tourism demand forecasting system.
5. To provide recommendations and suggestions for decision-makers in both the public and private sectors in Hong Kong.

All of the above objectives are basically aimed at improving the forecasting performance of existing tourism demand models. Tourism researchers and practitioners are interested in the accuracy of tourism demand forecasting for the following reasons. First, tourism demand is the foundation on which all tourism-related business decisions ultimately rest. Companies such as airlines, tour operators, hotels, cruise ship lines, and recreation facility providers are interested in the size and level of tourism demand for their products. The success of many businesses depends largely or totally on the state of tourism demand, and ultimate management failure is quite often due to the failure to meet market demand. Because of the key role of demand as a determinant of business profitability, estimates of expected future demand constitute a very important element in all planning activities. It is clear that accurate forecasts of tourism demand are essential for the efficient planning of tourism-related businesses, particularly given the perishable nature of tourism products. Second, tourism investment, especially investment in destination infrastructures such as airports, highways, and rail links, requires long-term financial commitments, and the sunk costs can be very high if investment projects fail to fulfil their design capacities. Therefore, the prediction of long-term demand for tourism related infrastructure often forms an important part of project appraisal (Wright & Ayton, 1987). Third, government macroeconomic policies largely depend on the relative importance of individual sectors within a destination. Hence, accurate forecasts of demand in the tourism sector of the economy will help destination governments to formulate and implement appropriate medium- to long-term tourism strategies.

In achieving the above objectives, this study targeted the following specific goals: (a) to compare and evaluate selected tourism forecasting techniques in terms of their forecasting performance (i.e. accuracy) based on proper error measures; (b) to structure and quantify experts' knowledge and expertise and integrate this knowledge and expertise into statistical forecasts in the context of tourism demand analysis; (c) to provide forecasts of inbound tourism demand from major source markets in Hong Kong for tourism decision-makers; and (d) to explore the judgmental forecasting behaviour of tourism practitioners and researchers.

To achieve the above goals, a mixed methods research approach was used to collect and analyse the quantitative and qualitative data of this study. The sequential explanatory strategy− the most straightforward of the six major mixed methods approaches summarized by Creswell (2009) − was adopted to utilize the combined strengths of qualitative and quantitative approaches. This strategy offers the best possible approach to build a deeper insight into the current topic which has little previous evidence to guide the research. In the first phase of the study, the quantitative techniques comprehensive evaluated the effectiveness of integrating statistical and judgmental forecasts, and this helped to address the first three specific goals ((a)-(c)). In the second phase, the follow-up in-depth interviews were conducted with a view to understanding the empirical findings obtained from the first phase as well as addressing the fourth specific goal.

1.4 Contributions of this Thesis

Apart from the impact of economic factors, the forecast accuracy of tourism demand models is also heavily influenced by such factors as policies, promotions, and planning activities. Typically, these external factors cannot be considered in statistical forecasting models as they often occur outside the forecasting period. In such cases, judgmental forecasting methods could be used to improve the forecasting performance of the statistical models by considering these factors.

In light of the complementary strengths and weaknesses of the two types of forecasts discussed previously and the mixed empirical evidence on their comparative performance, it is very likely that the integration of judgmental and statistical forecasts will contribute more to the improvement of forecast accuracy than either type could do on its own (Blattberg & Hoch, 1990). A large proportion of the recent forecasting literature has provided evidence, either through experiments or real data sets, to support this view (e.g. Goodwin, 2000a, 2000b, 2002).

The main aim of this study, therefore, is to improve the forecast accuracy of the tourism demand models developed by the investigators by incorporating judgmental forecasts into statistical forecasts. It is expected that the integration of these two types of forecasts will reduce the risk of forecasting failure and provide more reliable information for policymakers in the Hong Kong tourism industry. The theoretical and practical contributions of the study are discussed below.

1.4.1 Theoretical contributions

The majority of the published tourism demand forecasting studies have focused on the statistical (time series and econometric) approaches to forecasting, with very limited attention being paid to judgmental forecasting approaches in the tourism forecasting field. This study contributes to the tourism forecasting literature by providing experimental and empirical evidence on the efficiency of integrating judgmental and statistical forecasts with a particular focus on judgmentally adjusting statistical forecasts with the use of a Web-based forecasting support system.

One contribution of this study is to build up a systematic integration 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 to structure experts’ knowledge and quantify managerial intuition, (c) measures statistical and judgmentally adjusted forecasts using formal measures of accuracy, (d) evaluates different types of adjustments and explores the relationship between adjustments and forecast accuracy, and (e) explores the reasons for bias and inefficiency. 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. In addition, this study examines the applicability of an integrated framework in tourism demand forecasting under different circumstances. Last but not least, the use of this integrated framework is not only limited to the tourism industry: Any business corporation or government organization could use this framework to tackle the main issues associated with judgmental forecasting tasks.

1.4.2 Practical contributions

The findings of this study can be applied in the following two areas.

First, it is suggested that tourism managers and forecasters can systematically integrate managers' predictions into statistical forecasts, particularly short-term forecasting tasks under conditions of high uncertainty. Taking Hong Kong as an example, the high volatility exhibited in the demand data and the relatively high demand elasticities make formulating models to capture these features more difficult. This is largely because statistical methods are unable to cope with special events or new circumstances in the forecasting environment. An integration of statistical and judgmental forecasting methods, therefore, may prove to be useful in capturing the variability inherent in the tourism data.

From a practitioner's point of view, the means of incorporating a qualitative input into a formal forecasting system is of no little importance (Mathews & Diamantopoulos, 1986). It seems highly likely that a manager will reject at face value a forecast produced by a quantitative forecasting model, especially if the forecast is inconsistent with his/her own estimates. Therefore, information on the effects of judgmental adjustments would be of considerable use to a tourism manager in his/her use of a forecasting system. This study has been designed to examine tourism demand forecasts in a real world situation, and to the best of our knowledge, it is the first empirical attempt to directly address this issue. Specifically, the study aims to determine the extent to which managerial adjustment of arrivals forecasts results in superior or inferior projections by investigating the relative performance of arrivals forecasts before and after adjustment.

1.5 Structure of this Thesis

This thesis consists of six chapters, references, and appendixes (see Figure 1.3). Chapter 1 introduces the research background, problem statement, research objectives, and potential contributions of the present study, and the remainder of the thesis is structured as follows.

Chapter 2 provides an extensive literature review, the purpose of which is to address some theoretical and methodological gaps in the integration of judgmental and statistical forecasting methods/models. The chapter starts with a summary of previous findings on using different forecasting techniques, followed by a survey of empirical studies emphasizing the role and validity of incorporating judgment in the forecasting process. The strengths and drawbacks of quantitative and judgmental forecasting methods are compared by reviewing the empirical studies. Different types of methodologies to integrate judgmental and statistical forecasts are summarized. The literature review pays particular attention to the judgmental forecasting studies in the tourism research.

Next, Chapter 3 presents the research methodology and data analysis. This chapter includes a description of the data, variables, and econometric models of this study. After introducing the judgmental adjustment procedure within the Hong Kong Tourism Demand Forecasting System (HKTDFS), the chapter discusses the justifications for using in-depth interviews. A group of error measures and statistical tests are used to test the proposed hypotheses. Chapter 4 describes the results of the pilot study, which was carried out with the involvement of a group of postgraduate students and research staff. Chapter 5 presents the key findings from the Delphi survey followed by the hypothesis testing results. This chapter also summarizes the findings from the in-depth interviews, which were conducted to explore the possible causes of inaccurate, biased, and inefficient forecasts. Chapter 6 concludes the study by summarizing the key findings and addressing the limitations of the study and offering suggestions for the future research agenda.

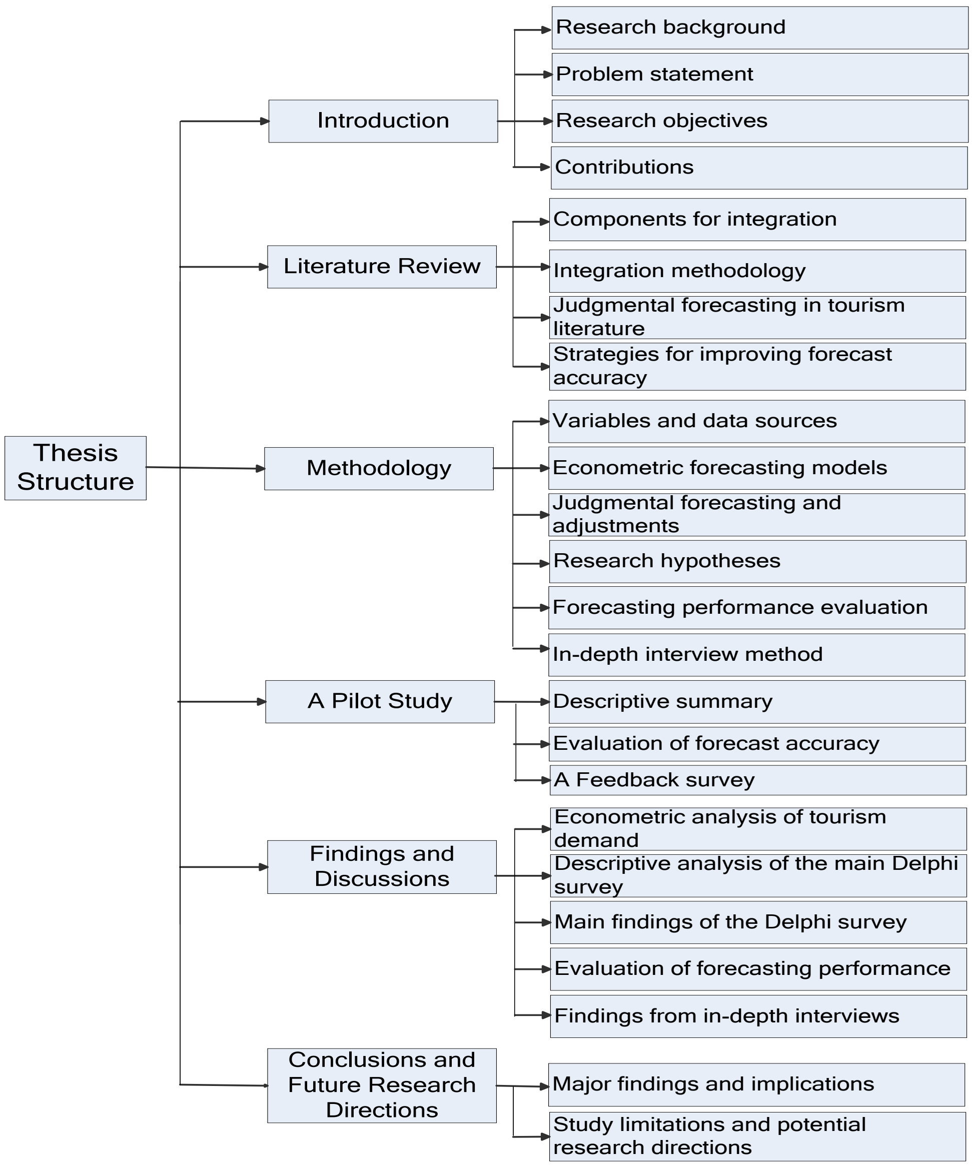


Figure 1.3 Structure of the thesis

1.6 Chapter Summary

Tourism demand forecasting is a widely researched topic as it has important policy implications for tourism planners, policymakers, and business executives. 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. This study aims to contribute to the existing literature by systematically evaluating the forecasting performance achieved by integrating statistical and judgmental forecasts using tourism demand data in Hong Kong.

This chapter started with a brief introduction to background information on the Hong Kong tourism industry. The problem statement was formulated by addressing the need to integrate tourism forecasters' judgments into statistical forecasts in tourism demand forecasting. The characteristics of tourism demand were described and discussed to support the existence of the problem and the need for integration. A brief overview of forecasting approaches was provided to illustrate recent developments in forecasting methodology. The research objectives and the contributions of the study were presented at the end of the chapter.

Chapter 2 : Literature Review

2.1 Introduction

The unprecedented growth in tourism across the world over the past five decades has generated considerable research interest from both industry practitioners and academic researchers. As an important area of tourism research, tourism forecasting has attracted much attention from both practitioners and academics, with impressive increases in publications since the 1960s. An overwhelming portion of these publications has been oriented toward quantitative approaches, with several reviews seeking to gain a better understanding of methodological developments and the empirical evidence on quantitative forecasting methods (Li, Song, & Witt, 2005; Song & Hyndman, 2011; Song & Li, 2008).

In the past four decades, the majority of the quantitative forecasting studies in the tourism field have focused on time-series methods and static regression models. Applications of the autoregressive integrated moving average (ARIMA) model proposed by Box and Jenkins (1976) and its extensions − such as seasonal ARIMA (SARIMA), multivariate ARIMA (MARIMA), and the generalized autoregressive conditional heteroscedastic (GARCH) model − have dominated time-series modelling and forecasting research. Naive 1 (or no-change) model, Naive 2 (or constant-growth-rate) model, exponential smoothing models, and simple autoregressive models have also been frequently used, and they are often used as benchmark models for accuracy evaluation (Song & Li, 2008).

Since the early 1990s, significant advances have been made in the application of econometric techniques in tourism forecasting (Li, 2009; Li, Song, & Witt, 2005;

Song & Guo, 2008; Song, Witt, Wong, & Wu, 2009). Advanced econometric forecasting techniques, such as the autoregressive distributed lag model (ARDL), the vector autoregressive (VAR) model, the error correction model (ECM), the time varying parameter (TVP) model, and almost ideal demand systems (AIDS), have been developed and employed in modelling and forecasting tourism demand (Li, 2009; Li, Song, & Witt, 2005; Song & Guo, 2008; Song et al., 2009). Many empirical studies have shown that advanced econometric approaches such as the TVP and ECM methods tend to generate more accurate tourism forecasts than traditional forecasting methods, but this finding cannot be extended to all individual cases (Song, Witt, & Jensen, 2003).

Some consensus has been reached on the fact that no single model or method outperforms others on all occasions (Song & Li, 2008; Witt & Witt, 1995). Rather, the most appropriate model for a forecasting task should be determined by environment-specific conditions. Careful decisions must be made when a number of alternatives exist so that an appropriate forecasting method can be selected and adopted for the specific situation considered.

Despite the overwhelming dominance of quantitative methods, it is of great importance to pay attention to the judgmental (or qualitative) methods. The term “judgmental forecasting” is more often associated with forecasts that are made based entirely on the basis of judgment or with judgmental adjustments to statistical forecasts (Wright & Goodwin, 1998). Without a doubt, human judgment is never isolated from the forecasting process (Clemen, 1989; Makridakis et al., 1982). Even for those forecasts generated by sophisticated statistical models, judgment has to be relied on in the selection of a particular forecasting method, functional form, dependent variables and regressors, and data sets (Goodwin, 2002). Judgmental forecasting methods depend on the accumulated experience of individual experts or groups of experts to make projections about the event concerned. Under certain circumstances, such as (a) insufficient historical data, (b) the unreliability or invalidity of the available time series, (c) rapid changes in the macroenvironment, (d) expectations of major disturbances, and (e) a desire for long-term forecasts, judgmental forecasting methods are likely to generate more accurate forecasts than simple statistical methods (Frechtling, 2001).

The contribution of academic research to understanding and assessing the role of experts' judgment in tourism forecast accuracy, however, remains limited. Witt and Witt (1995) explained that qualitative forecasting methods lack popularity because "they are just standard applications" (p. 460) from a methodological perspective. Among the judgmental methods, Delphi and scenario writing have been two of the most popular techniques (Calantone, Benedetto, & Bojanic, 1987; Witt & Witt, 1995). Some review articles reported in Tisdell (2000), such as Archer (1980), Calantone, Benedetto, and Bojanic (1987), Witt and Witt (1995), and van Doorn (1982, 1984), briefly summarized the empirical applications of judgmental forecasting techniques in tourism (though mainly focusing on Delphi and scenario projections).

Based on the vast empirical evidence from the general forecasting literature, the expectation is that integrating judgmental and statistical methods will improve accuracy in tourism demand forecasting, but to date, little effort has been put into examining the effects of integrating these two types of approaches in the tourism context. This study is intended to fill this gap. To discover the most appropriate approach that can achieve the research objectives of this study, the starting point is to provide a comprehensive review of relevant studies. To produce this review, an extensive literature search based on various databases, such as the Social Science Citation Index (SSCI), Science Citation Index Expanded (1970+), Google Scholar, ScienceDirect, and EBSCOhost, and citations of published articles, was conducted; citations from identified articles were also traced.

The rest of this chapter is structured as follows: Section 2.2 explores the role and validity of judgment in statistical forecasting methods and compares the forecasting performance of statistical methods with that of judgmental methods; Section 2.3 introduces five approaches to facilitating the integration process; Section 2.4 reviews the judgmental forecasting studies in the tourism literature; Section 2.5 summarizes a number of strategies for improving the accuracy of judgmental and statistical forecasting methods; and Section 2.6 concludes the chapter.

2.2 Components for Integration: Judgmental and Quantitative Methods

At least three methods are available for incorporating judgments into quantitative forecasts. First, researchers can rely on decisions based on human judgment about what data are relevant and required in the forecasting tasks. Such task requires forecasters to decide what types of data to use in producing forecasts and involves making decisions on matters such as data sources (i.e. survey data or public data), data frequency (i.e. daily, monthly, quarterly or annual data), and forecasting variables (i.e. tourist expenditure or tourist arrivals). Such decisions vary according to the specific problems being addressed. Once a time series is chosen, forecasters should first decide whether it should be forecast directly or decomposed made. Second, the choice of selecting the best forecasting methods is determined by the forecasters. Data availability will determine what methods to employ and who should conduct the analysis; for example, if only a historical time series is available and the forecasting task is to predict future values, extrapolation can be used. Nevertheless, unless the causal factors of the variables to be forecast are available, regression methods are probably a better option. Third, forecasters' technical knowledge and domain knowledge can be incorporated into the forecasts. Forecasters can use their expertise to select the variables of interest, directly make forecasts, revise time-series observations, adjust unusual data points, make projections about the future effects of causal variables, and facilitate model building in the quantitative forecasting process. Alternatively, decision support systems can be used to structure domain knowledge. Before discussing when and how to integrate judgment into quantitative forecasts, it is important to gain a good understanding of what role judgment can play in the forecasting process and then compare the performance of judgmental forecasts and quantitative forecasts.

2.2.1 The role and validity of judgment

Judgment has been studied for many years by psychologists interested in human reasoning and decision-making (Ghalia & Wang, 2000; Wright & Ayton, 1987). Most studies have been conducted from the perspective of "decision theory", a term originating from statistics and economics. Decision theory proposes that two independent types of information are involved when making good decisions: one is the subjective values or utilities attached to the outcomes of events at some time in the future, while the other is the subjective probabilities attached to those events occurring (Wright & Ayton, 1987).

All forecasting involves human judgment, and even sophisticated statistical methods such as the Box-Jenkins time-series models (which require judgment in the model identification process) and multiple regression models (which require judgment in the variable selection and model identification process) rely heavily on judgment. There is much evidence, supported by numerous surveys from corporate forecasting practices, that these judgmental methods have been most frequently involved to facilitate strategic decision-making in many business activities.

Based on a questionnaire survey of 175 businesspeople, Dalrymple (1975) reported that executives used the jury of executive opinion and sales force composite methods significantly more than statistical forecasting procedures. Dalrymple (1975) also concluded that the judgmental forecasting results yielded an average forecast error of 7 per cent. Dalrymple (1987) surveyed 860 companies in the USA and found that subjective techniques remained popular and that the Naive model was used by a surprising large number of firms for short- and medium-range predictions. Similarly, among 52 surveyed manufacturing firms, Rothe (1978) found that 50 of them used judgmental methods in one form or another and that opinion-based forecasting methods were the most popular forecasting method. Based on a survey of 500 of the world's largest corporations, Klein and Linneman (1984) found that the overwhelming majority of corporate planners recognized the limitations of using purely statistical techniques. They also found that a variety of judgmental (or speculative, conjectural) techniques had became accessible in direct response to the turbulent environmental conditions worldwide. Sanders and Manrodt (1994) surveyed 500 corporations in the USA and found that managers and practitioners relied heavily on judgmental forecasting methods and that the level of using statistical forecasting methods had not increased even among managers who were more familiar with these approaches.

In terms of the application of econometric models, substantial subjective components are incorporated into the mathematical forecasts, reflecting the forecasters’ own personal projections about the future (Winklhofer, Diamantopoulos, & Witt, 1996). Actually, experienced managers and forecasters often manually adjust quantitative forecasts as they have their own perceptions of future trends based on the events occurring in the market and their prior expertise and industry experience.

Sanders (1975, cited in Makridakis & Wheelwright, 1979) found that the majority of the manufacturing and service companies he investigated “always” or “frequently” adjusted their quantitative forecasts to add external information about the environment, the product, and past experience. Sanders and Manrodt’s (1994) study revealed that professionals often input their subjective judgments to adjust forecasts when adopting quantitative forecasting approaches. Walker and McClelland (1991) reported that the annual forecasts of the companies in their study were judgmentally adjusted to incorporate advertising and sales promotion activities and sales forecasts were generally adjusted by the vice president: the underlying reasons for making such adjustments ranged from “gut feeling about the future sales trends” to “more specific anticipated efforts of planned selling and marking programs” (Armstrong, 1985, p. 80). Thus, there is an argument about how many quantitative forecasts should be included in order to obtain better forecasts and how much the judgmental efforts of forecasters will affect the results.

(1) The quality of judgment research

No forecasts can perfectly reflect reality. There are many types of errors derived from judgmental forecasts, among which **bias** and **anchoring** are the two most serious issues (Armstrong, 1985). **Bias** arises from the researcher and from the situation, but the most serious form is caused by the judge. Judges have preconceived notions about the forecasting problems that can greatly influence their forecasts. One form of this bias is called “optimism” (Armstrong, 1985). Ogburn (1934) and MacGregor (1938) found that judgments were strongly influenced by biases such as favouring a desired outcome (optimism bias).

Anchoring is the tendency to start with an answer when making a forecast (Armstrong, 1985). It is a type of bias that is developed by judgmental forecasters when starting with an initial forecast as an anchor and adjusting from it to obtain the final forecast (Goodwin, 2005; Makridakis, Wheelwright, & Hyndman, 1998). The problem associated with anchoring is that the adjustments are usually too small and so the final forecasts are too close to the anchor, which may explain the widely observed tendency of judgmental forecasters to underestimate upward trends (Goodwin, 2005; Stekler, 2007). One common type of anchoring is *conservatism* which assumes that the future will resemble the past and so there will be no abrupt changes. Conservatism often underpredicts the amount of change; for example, Eggleton (1982) found that judgmental forecasts were more conservative than extrapolation forecasts.

(2) The role of domain knowledge and contextual information

The existing literature has not provided clear definitions of “contextual information” and “contextual knowledge” as contextual information and the forecaster’s experience are not clearly separated. Webby and O’Connor (1996) defined “contextual information” as “information, other than the time series and general experience, which helps in the explanation, interpretation, and anticipation of time series behaviour” (p. 97). This definition encompasses the labels (e.g. information that a series represents costs or sales) defined earlier by Goodwin and

Wright (1993).

Webby and O’Connor (2001) further distinguished between contextual information and domain knowledge. The former refers to the information available in the forecasting environment, while the latter refers to the knowledge of the forecaster. In other words, contextual information is an attribute of the forecasting environment, whereas domain knowledge is an attribute of the forecaster. Webby and O’Connor (2001) argued that domain knowledge is the result of applying human interpretations to contextual or environmental information. The quality of domain knowledge depends on forecasters’ ability to acquire an appropriate understanding of contextual information. However, contextual information may not necessarily produce corresponding domain knowledge.

According to the findings of four review studies (Goodwin & Wright, 1993, 1994; Lawrence et al., 2006; Webby & O’Connor, 1996), contextual information is one of the key determinants of judgmental methods’ superiority over statistical models. When time-series data are unstable, contextual information is particularly favourable, presumably due to the greater number of discontinuities that can only be explained by human judgment (Goodwin & Wright, 1993, 1994; Lawrence et al., 2006; Webby & O’Connor, 1996). In addition to contextual information, other factors such as trend, instability, historical data points, length of forecasting horizon, and high seasonality with the presence of low instability may also affect the accuracy of judgmental forecasts, but perhaps only subject to the absence of contextual information.

2.2.2 Statistical versus judgmental forecasting methods

A number of studies have indicated that forecasters and practitioners choose their forecasting techniques without considering the particular forecasting situation (Armstrong, 2001e). Makridakis and Wheelwright (cited in Wright & Ayton, 1986) showed that forecasters tend to "concentrate on well-behaved situations that can be forecast with standard methodologies" and "ignore the rapidly changing situation for which management may most want forecasts" (p. 421). However, even where judgmental forecasts are less accurate than statistical forecasts, managers may persist in their own judgments that are more acceptable to them (Rothe, 1978; Winklhofer & Diamantopoulos, 2002).

Many arguments have been put forward on comparing the advantages and disadvantages of these two forecasting methods (see Table 2.1). Statistical forecasting methods are capable of filtering regular time-series patterns from "noisy" data, while judgmental forecasting tends to overreact to random movements in series (O'Connor et al., 1993, cited in Goodwin, 2000a). Statistical methods can make efficient use of historical data, particularly when a large amount of historical data is involved. The greater the availability of the data, the more efficiently the statistical methods can capture the patterns of the series. On the other hand, judgmental methods incorporate expert opinions on likely outcomes and possible alternative scenarios and are used in the absence of reliable historical data. However, one problem associated with such a forecasting approach is that it tends to be subjective and the assumptions upon which the forecasts are produced are not always justified or even made explicit. In addition, judgmental forecasting is subject to cognitive biases when a large volume of information is involved (Hogarth, 1985).

Table 2.1 A comparison of statistical versus judgmental methods

Situation I: Statistical forecasts better than judgmental forecasts	
Experts	*Models*
Suffer from human bias	Unbiased
Suffer from overconfidence	Take base-rates into account
Influenced by organizational politics	Immune to social pressures or consensus
Emotional problems (e.g. get tired, bored)	No emotional problems
Inconsistently integrate evidence	Optimally weigh the evidence
Situation II: Judgmental forecasts better than statistical forecasts	
Experts	*Models*
Raise questions and explore reasons	Only respond to forecasters' inputs
Identify new variables	Could not identify
Diagnose and predict	Predict only
Proficient in attribute valuation, provide subjective evaluations of variables that are difficult to measure objectively	Only proficient in dealing with large amounts of quantified information
Consistent, but rigid	Inconsistent, but flexible
Highly organized, domain-specific knowledge (i.e. may recognize and interpret abnormal cases with "broken leg" cues)	Could not incorporate up-to-date knowledge of changes and events occurring in the environment that can affect the variable being forecast

Source: Adapted from Blattberg and Hoch (1990, pp. 889-890).

Qualitative approaches may be particularly advantageous for medium- and long-range situations such as formulating strategy, developing new products and technologies, and making long-range plans (Makridakis, Wheelwright, & Hyndman, 1998). It is, however, difficult to measure the usefulness of judgmental forecasts that are used mainly to provide hints, aid decision makers and planners, and supplement quantitative forecasts rather than to provide specific numerical forecasts.

Statistical forecasts can deal with instability better than judgmental forecasts in situations where contextual information is absent (O'Connor, Remus, & Griggs, 1993; Webby & O'Connor, 1996). However, the opposite is true if contextual information is available for consideration (Sanders & Ritzman, 1992). Numerous empirical studies have been carried out to compare judgmental and statistical forecasts. For example, Goodwin and Wright (1993) found that comparative forecasting performance depends on various factors, such as the nature of the time series (e.g. trend, seasonality, noise, instability and forecasting horizon) and situational characteristics. Goodwin (2002) also contended that judgmental forecasts are expected to be superior to statistical methods in dealing with trending series. Other researchers have argued that the performance of statistical forecasts depends on the selection of the statistical techniques used for comparison.

Early comparisons of judgmental forecasting and statistical forecasting methods used artificial data and reached equivocal conclusions regarding the relative accuracy of the two methods (Eggleton, 1982; Lawrence et al., 2006). The first large-scale comparison of the accuracy of judgmental forecasts and statistical forecasts using real life data was conducted by Lawrence, Edmundson, and O'Connor (1986) using what is now known as the M1 forecasting competition (Makridakis et al., 1982). Lawrence et al.'s (1985) study concluded that "judgmental forecasting … [is] at least as accurate as statistical techniques, while in a number of subgroupings of the time series a judgmental technique was the most accurate" (p. 34). Lawrence, Edmundson, and O'Connor (1986) stated that combining judgmental forecasts and statistical forecasts produces higher accuracy than using two statistical forecasts because judgmental forecasts are less correlated with statistical forecasts. Lobo (1992), Sanders (1992), and Makridakis et al. (1993) provided further empirical evidence to support this finding.

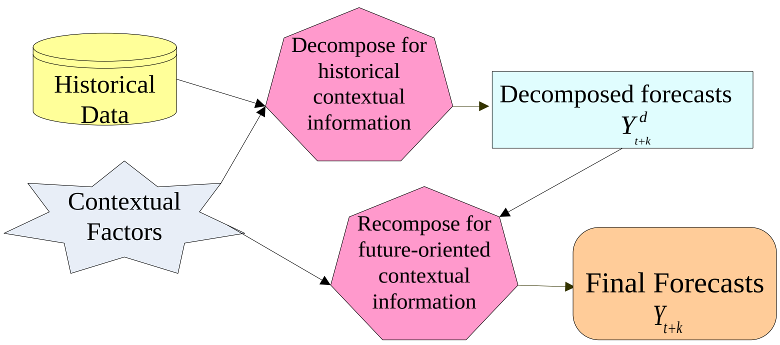
Statistical and judgmental forecasting methods have unique forecasting characteristics: the former are too consistent, while the latter are too flexible. They are both substitutes (because both take into account much of the same decision relevant information) and complements (because where one decision input is weak the other is stronger and vice versa) (Blattberg & Hoch, 1990). As judgmental and statistical methods each have their unique strengths and weaknesses (see Table 2.1), it makes sense to bring together the advantages of each method and offset their shortcomings to improve forecast accuracy.

2.3 Integration Methodology

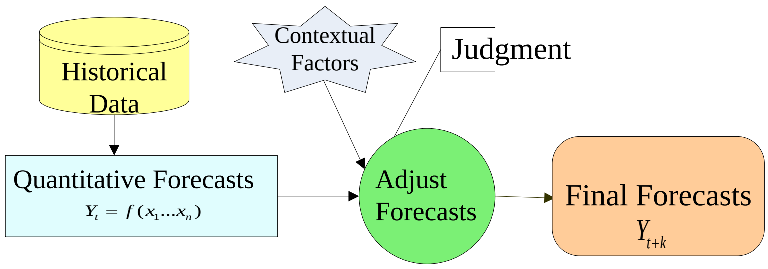
A number of methodologies for integrating judgmental and statistical forecasts have been proposed. However, the type of method used for integration varies according to the researchers' research purposes. The first method is to judgmentally adjust the statistical forecasts. The second method is to forego statistical techniques entirely: forecasters simply use additional information and the time-series data as the basis for making judgmental forecasts. The third method is to use judgment as an input into statistical methods, and the fourth is to combine independent judgmental and statistical forecasts. Alternatively, decomposing the forecasting tasks into several elements also complements the integration process.

Webby and O'Connor (1996) identified the four most commonly used approaches to integrating judgmental (subjective) and statistical (objective) forecasts: *model building*, *forecast combination*, *judgmental adjustment*, and *judgmental decomposition*. These approaches were reviewed again by Lawrence et al. (2006) with the provision of more empirical evidence. Starting from the domain knowledge, Armstrong and Collopy (1998) presented five procedures to facilitate the interaction of judgment with structured forecasting methods: (a) revise judgmental forecasts, (b) combine forecasts, (c) revise extrapolation forecasts, (d) use rule-based forecasts, and (e) use econometric forecasts. Further elaboration of these procedures can be found in Armstrong (2001d). However, only a few studies have considered the role of human behaviour and organizational context as input factors in the success of methodology implementation. One attempt to fill this gap was made by Sanders and Ritzman (2004), who considered human factors (e.g. ownership) and organizational factors (e.g. location of final forecast generated within the organization) in the integration process.

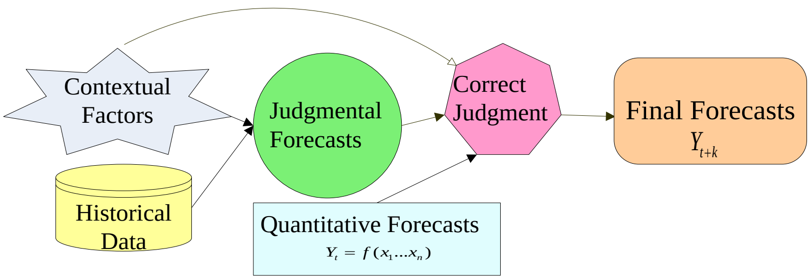
Based on a comparison of previous reviews on judgmental forecasting and statistical forecasting (Armstrong & Collopy, 1998; Bunn & Wright, 1991; Goodwin, 2000b; Goodwin & Wright, 1993; Lawrence et al., 2006; Webby & O'Connor, 1996), five major methods for integration were identified: (a) decomposition, (b) the adjustment of statistical forecasts, (c) the quantitative correction of judgmental forecasts, (d) the combination of judgmental and statistical forecasts, and (e) model building. Figure 2.1 presents a diagrammatic classification of the five approaches involved in the forecasting process in increasing order of objectivity; it also demonstrates how the judgmental and quantitative forecasting processes may interact in each method.



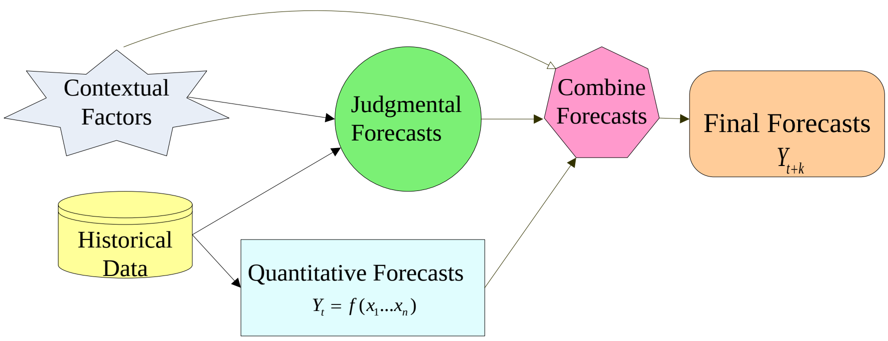
Method 1: Decomposition



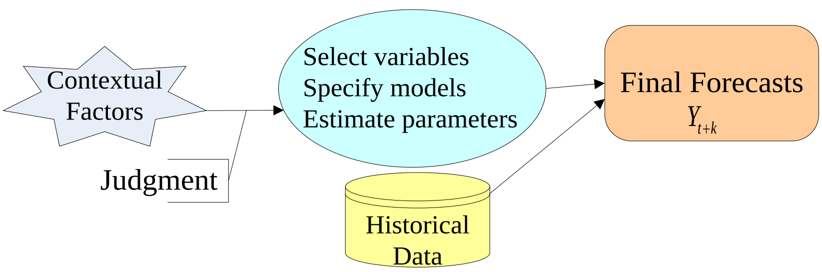
Method 2: Adjustment of quantitative forecasts



Method 3: Quantitative correction of judgmental forecasts



Method 4: Combination of judgmental and quantitative forecasts



Method 5: Model building

Figure 2.1 Integrating subjective and objective forecasts

Sources: Adapted from Sanders and Ritzman (2004), and Webby and O'Connor (1996).

2.3.1 Model building

The formulation of a statistical model requires the input of many aspects of judgment (Bunn & Wright, 1991). *Model building* is an approach that integrates judgment into every stage of the statistical forecasting process, including selecting the variables, deciding the functional form, building models, estimating parameters, and conducting data analysis. This method requires considerable cross-functional integration and information sharing; thus it is considered to be the most effective and most objective integration method (Bunn & Wright, 1991).

All of the various types of model, such as exponential smoothing, regression, the ARMA model, and decomposition, have different judgmental problems and incorporate additional subjective information with varying degrees of facility (Bunn & Wright, 1991). Armstrong and Collopy (1998) defined integration as an econometric model where judgments are used to identify a model and regression is used to estimate the coefficients of this model. They contended that econometric models provide the most highly structured approach to integrating judgments. Research has shown that when judgment is based on good domain knowledge, econometric models can exhibit higher forecast accuracy than alternative procedures when large changes are expected (Armstrong, 1985; Fildes, 1985).

2.3.2 Forecast combination

(1) The framework for a combination of forecasts

One of the underlying justifications for the concept of combined forecasts is that no one forecasting method is perfect enough to fully capture reality. The best forecasting method is usually defined as the method with the lowest forecast errors. Different error measures, such as absolute error, squared error, and percentage error, may result in different conclusions. However, it is extremely rare to find a forecasting method that is superior to other methods over all time horizons and time series. By combining different forecasting methods rather than seeking the best forecasting method, forecasts are created by integrating a set of hopefully good base forecasts.

The seminal works of Reid (1968) and Bates and Granger (1969) are reported in the general forecasting literature. Reid (1968) reported the results of an experiment to combine forecasts optimally on the basis of variances and covariances. Bates and Granger (1969) found that combined forecasts yield much lower forecast errors than individual forecasts. Clemen (1989) extensively reviewed the works on the development and applications of combined methods in various areas of forecasting before 1989 and argued that forecast accuracy could be substantially improved by combining individual forecasts. A large number of empirical and simulation studies have suggested that combination techniques can outperform the best constituent single individual forecasts (Armstrong, 2001a).

There are two major practical reasons for encouraging the use of combined forecasts. First, decision-makers can obtain information from a variety of sources in order to reduce or detect bias in the forecasts. Second, it may be relatively inexpensive to repeatedly use the same database to produce different sets of forecasts by using different forecasting methods. These databases may have been established in separate areas within an organization, either from external or internal sources or both. In the abovementioned cases, a decision should be made to determine the best forecasting method or to pool different forecasts into a final, combined forecast.

The term “combination” has two meanings. First, it refers to combining different methods to solve the forecasting problem being examined. Second, it is concerned with the relationship between the methods and the decision-making process itself. Flores and White (1988) proposed a research framework for analysing combined forecasts. This suggested framework is structured into two dimensions: selection of the base forecasts and selection of the combination method (Figure 2.2).

Base Forecasts	**Combination Technique**	
	Systematic	Intuitive
Quantitative	A	B
Judgmental	C	D
Both	E	F

Figure 2.2 Framework for combining forecasts

Source: Flores and White (1988, p. 97).

Figure 2.2 shows that there are three types of base forecasts in combined forecasting. The ‘quantitative’ category contains all of the quantitative forecasting methods, such as simple extrapolative methods (e.g. Naive methods, exponential smoothing methods), time-series methods, and econometric methods. The ‘judgmental’ category encompasses situations where base forecasts are produced by qualitative or subjective methods (e.g. expert opinions). The ‘both’ category is the integration of quantitative and judgmental forecasting methods, which reflects the current trend in the general forecasting research.

Before the 1990s, most of the research, such as Makridakis et al. (1982), Makridakis and Winkler (1983), Granger and Ramanathan (1984), Diebold and Lopez (1996), was concentrated on the systematic combination of quantitative forecasts (Cell A). Researchers have called for more research efforts into involving intuitive and systematic combinations between subjective forecasts (Bates & Granger, 1969; Lawrence, Edmundson, & O'Connor, 1986).

The studies in Cell C, such as Bunn (1975, 1981), Ashton and Ashton (1985), Bessler and Chamberlain (1987), and Hurley and Lior (2002), investigated the quantitative combination of subjective base forecasts. The intuitive combination of either objective or subjective forecasts (Cells B and D) has been relatively less explored. One such study was conducted by Lawrence et al. (1986), who reported on the combination of judgmental and statistical forecasts in both systematic and intuitive form. Another study was provided by Flores and White (1989) (Cell B), who conducted an experiment to evaluate the subjective and objective combination of quantitative forecasts.

The literature has suggested that there is no single 'best' combination method, but one lasting conclusion is that almost "any combination of forecasts proves more accurate than the single inputs" (Blattberg & Hoch, 1990, p. 889). The existing forecasting literature has considered 'model-model' (Cell A) and 'expert-expert' (Cell D) combinations. However, an increasing amount of research effort has been devoted to the 'model-expert' combination (Cells E to F). For example, Armstrong and Lusk (1983) surveyed a group of experts and identified a need to integrate judgment into extrapolation. Bunn and Wright (1991) concluded that judgmental and statistical forecasting methods should be integrated. Examples of studies which falls into Cells E and F include Lawrence et al. (1986), Blattberg and Hoch (1990), Lobo and Nair (1990), Lobo (1991), and Sanders and Ritzma (1995).

(2) Integration of judgmental and statistical methods

There has been a growing recognition of the value of integrating statistical forecasts with judgment, and this type of research has received much support in the literature to date. The effectiveness of the integration between judgmental and statistical forecasting methods has been documented in many studies (Clemen, 1989; Goodwin, 2002; Lawrence et al., 2006; Webby & O'Connor, 1996). The empirical evidence has demonstrated its significant contribution to accuracy improvement in comparison to individual forecasts, and one general observation from the existing literature is that integration improves forecast accuracy because the constituent forecasts can provide different aspects of the information available for producing forecasts (Clemen, 1989). Lawrence, Edmundson, and O'Connor (1985) examined the effectiveness of combined forecasts for real-life economic time series with different levels of forecasting difficulty and seasonality and found that according to the mean absolute percentage error (MAPE), a mechanical integration of judgmental and statistical forecasts is superior to individual forecasts. Weinberg (1986) found that economic forecasts (MAPE, 31.2%) alone were more accurate than managers' forecasts (34.8%), but a mechanical combination of the two forecasts was superior to both (28.9%).

The integration of judgmental and statistical forecasts can be made either objectively or subjectively in light of the specific contextual information provided. The two approaches to integrating judgmental and statistical forecasting methods are 'mechanical integration' (or 'mechanical combination') and 'voluntary integration' (Sanders & Ritzman, 1990). Mechanical integration has at least three advantages over voluntary integration: (a) it is more objective and avoids the introduction of biases or political manipulation, (b) it is easier to disclose fully the process that generates them, and (c) it tends to be more accurate because it uses knowledge in a more efficient manner (Armstrong & Collopy, 1998).

Armstrong (2001a) showed that researchers preferred the mechanical integration of judgmental and statistical forecasts over voluntary integration (judgmental adjustment) as the latter is more subject to the negative effects of the judgment. The constituent forecasts are generated in parallel with the final forecasts obtained from the mathematical integration of the two (Sanders & Ritzman, 2004). Furthermore, the final forecasts represent a pooling of information upon which the constituent forecasts are based.

A starting point for mechanical integration is the simple average of judgmental and statistical forecasts that are made independently (Goodwin, 2002), while combination, correction for bias, and bootstrapping are the typical methods used. When only time-series information is available, voluntary integration or judgmental adjustments of statistical forecasts are used.

Makridakis and Winkler (1983) examined the forecasts of 1,001 series and found that the accuracy of the mechanically combined forecasts depended on both the number of methods in the averaging process and the specific methods being combined. They also found that the variability associated with the choice of methods was reduced as more methods were included. After combining seven extrapolations for 103 consumer products, Schnaar (1986) found that for one-year-ahead forecasts, combined forecasts led to a corresponding reduction of forecast errors by 1.8 per cent, and a more significant reduction of forecast errors (7.5%) when the five-year-ahead forecasts were concerned. Compared to the typical component forecasts, combined forecasts are always much more accurate, with error reductions in the MAPE exceeding 12 per cent (Armstrong, 2001a). Under ideal conditions (high uncertainty and combining many valid forecasts), the error reduction can even exceed 20 per cent and combined forecasts are often more accurate than the best individual forecasts.

The incorporation of judgmental forecasts based on contextual knowledge into combined forecasts has been found to improve forecast accuracy over individual statistical and judgmental forecasts, especially where the time-series data have a high degree of variability (Sanders & Ritzman, 1995). Sanders & Ritzman's (1995) findings also indicated a linear relationship between the amount of contextual knowledge needed and data variability. Lim and O'Connor (1996) also suggested that one should incorporate any contextual knowledge into an independent judgmental forecast and mechanically combine it with other forecasts. By contrast, Harvey and Harries (2004) argued that forecasters should not include judgments in the combination because they are likely to over-weigh their own judgments. In light of the above concerns, mechanical integration has been recommended by some researchers, such as Lim and O'Connor (1995), Goodwin and Fildes (1999), and Goodwin (2000a).

On the basis of an extensive review of over 200 empirical studies on combined forecasts, Clemen (1989) found that a mechanical combination helps to eliminate bias and enables a full disclosure of the forecasting process. There is abundant evidence to show that a mechanical combination of judgmental and statistical forecasts improves forecast accuracy or is likely to result in accuracy improvement; examples included the studies of Blattberg and Hoch (1990), Goodwin (2000a, 2000b, 2002), and Lobo and Nair (1990). In addition, Hibon and Evgeniou (2005) showed that the advantage of combined forecasting is not that the best possible combinations perform better than the best possible individual component forecasts, but rather that in terms of reducing forecasting risks, it is better to combine forecasts than to select an individual forecasting method.

This integration approach is especially useful when the component methods differ substantially from one another. For example, Blattberg and Hoch (1990) demonstrated that when managers' judgmental forecasts are combined with forecasts from quantitative models, the result is more accurate than either of the individual forecasts where accuracy is measured by correlation coefficients. One influencing factor that affects the forecast accuracy of combined forecasts is the correlation between the errors of the forecasts in the combination (Goodwin, 2002). Goodwin (2002) disclosed that a combination is likely to be less effective when the correlation between the forecast errors is high as the second forecast does not bring much new information into the combination; in other words, mechanical combination is most effective when there are greater forecasting divergences (e.g. when the forecasts generated by different models are negatively correlated).

A mechanical combination is most effective when the component forecasts are not correlated and can bring different information to the forecasting process. A number of methods have been proposed to estimate the constituent weights when mechanically combining two independent forecasts. The simplest method is to apply an equally weighted average of individual methods. Armstrong and Collopy (1998) recommended this as a starting point. However, under some conditions, an unequally weighted scheme may be more effective.

Numerous empirical studies have suggested that equally weighted averages are typically as accurate as any other weighting schemes lacking well-structured domain knowledge (Armstrong, 2001a). Hence, more weight should be given to formal quantitative methods when there are no major changes in the environment, but heavier weight should be placed on judgmental inputs when significant changes are expected (Armstrong & Collopy, 1998; Blattberg & Hoch, 1990; Clemen, 1989).

Lim and O'Connor (1995) found that there is an extremely strong tendency to place too heavy a weight on one's own forecasts rather than on statistical forecasts even when attention is drawn to the superior accuracy of the statistical forecasts. Likewise, Goodwin and Fildes (1999) found that judgmental forecasters carry out voluntary integration inefficiently. However, Fischer and Harvey (1999) found that when performance feedback relating to each of the individual forecasts available for combination is provided to forecasters, the accuracy of judgmental combination surpasses that of the simple averages.

Based on forecasters' objectives and the three properties of forecast errors, namely variance, asymmetry, and serial correlation, de Menezes, Bunn, and Taylor (2000) provided useful guidelines on selecting appropriate combining approaches when applying a mechanical combination. The properties of individual forecast errors could strongly influence the characteristics of combination errors; for example, when combining forecasts, forecasters should not only estimate the mean forecast error but also measure the asymmetry measures (e.g. mode, median) when analysing the forecasting results (de Menezes, Bunn, & Taylor, 2000).

2.3.3 Judgmental adjustment

The *judgmental adjustment* of statistical forecasts is a major competitive alternative to combining statistical and judgmental approaches. Numerous industry surveys have revealed that judgmentally adjusted statistical forecasting is a common practice. In a study of 96 corporations in the USA, Sanders and Manrodt (1994) found that 45 per cent of the respondents claimed that they always adjusted statistical forecasts and that 37 per cent did it sometimes. The main reason they gave for revising quantitative forecasts was to incorporate knowledge of the environment.

Similarly, Klassen and Flores (2001) surveyed 117 Canadian firms and found that senior management frequently revised the forecasts. They also found that 80 per cent of the respondents used computer-generated statistical forecasts and then judgmentally adjusted them, and this led to an average improvement in accuracy of 7.2 per cent.

A forecaster's goal in judgmentally adjusting a statistical forecast is to improve forecast accuracy by combining the relative strengths of statistical and judgmental methods (Armstrong, 2001a). Studies on the accuracy of judgmental adjustments, however, have reported equivocal results. Some researchers have found that judgmental adjustments improve forecast accuracy. Fildes and Goodwin (2007) found a median improvement in the absolute percentage error of about 7 per cent, which was slightly higher than the results (between 2.6% and 5%) reported in Fildes et al.'s (2006) study.

Some other 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 Naive method and the forecasting method in use) was high.

One major problem of using judgmental adjustment is that people read systematic patterns into the noise associated with a time series and this makes damaging adjustments to statistical forecasts (Lawrence et al., 2006; O'Connor, Remus, & Griggs, 1993). Most extrapolation methods cannot deal with discontinuities or pattern changes in the data. 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. Several studies have shown that under certain conditions, adjustments could improve the accuracy of statistical forecasts. 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 curtailed contextual information that is not included in the statistical method. In the first condition, Willemain (1991) suggested that the deficiency of a statistical method can be detected by comparing its accuracy relative to the Naive forecasts. Collopy and Armstrong (1992) found that forecasters who can identify the patterns in the data and incorporate contextual information in making judgmental adjustments to statistical forecasts can improve the accuracy of the forecasts. Sanders and Ritzman (2001) argued that statistical forecasts should be judgmentally adjusted in situations of high uncertainty. 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. Judgmental adjustments might also improve accuracy if the forecasters are able to make use of causal information that the statistical method had not used. 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.

The extant literature identifies two stages in the adjustment process: (a) a decision on whether a statistical forecast needs adjustment and (b) an estimate of the size of the adjustment that is required. However, most research has focused on the first stage and investigated ways in which forecasters can be discouraged from making gratuitous adjustments to statistical forecasts. Sanders and Ritzman (2001) suggested the following guidelines for judgmentally adjusting forecasts: (a) structure the judgmental adjustment process with either a computer-aided decision support system or paper and pencil; (b) document all judgmental adjustments made and periodically relate to forecast accuracy as this can enable forecasters to monitor their accuracy over time in order to evaluate forecasting performance as well as to discourage politically motivated biases; and (c) consider mechanically integrating judgmental and statistical forecasts rather than adjusting.

A substantial part of the research on judgmental adjustments has been conducted in experimental settings that may or may not be representative of an actual organizational setting. Goodwin and Wright (1993) identified 11 aspects in which an experimental study might potentially fail to represent an organizational setting where the statistical forecasts were judgmentally adjusted. Therefore, the results of experimental studies relevant to judgmental forecasts are usually not generalizable. In fact, the validity of the results obtained from experimental studies with a flawed methodology are questionable. Nevertheless, when undertaken under more realistic conditions, the value of judgmental adjustments becomes clear. Thus, researchers have been encouraged to conduct more studies in realistic conditions (Önkal & Gonul, 2005).

2.3.4 Quantitative correction of judgmental forecasts

In addition to combination, correction is another method of mechanical integration that has been proposed for situations where the forecasts are expressed as point estimates (Goodwin, 2000a). The combination of forecasts is derived by calculating a simple or weighted average of independent judgmental and statistical forecasts (Clemen, 1989). Correction methods involve the use of regression methods to forecast error in judgmental forecasts, and this predicted error can then be used to correct the judgmental forecasts.

Correcting judgmental forecasts for bias relies on judgment for forecast generation by performing a quantitative correction directly on to judgmental forecasts, which reduces the negative effects of judgment. Correction methods are most appropriate when the biases associated with judgmental forecasts are systematic and sustained (Lawrence, O'Connor, & Edmundson, 2000).

To date, this methodology has received less attention in the general forecasting literature. Arguably, however, correction, in its simplest form, is more convenient in that it does not require the identification, fitting, and testing of independent statistical methods in addition to the elicitation of judgmental forecasts (Goodwin, 2000a). A few studies have shown that this bias correction can result in large improvements in forecast accuracy (Ahlburg, 1984; Elgers, Lo, & Murray, 1995; Fildes, 1991; Goodwin, 1996; Sanders & Ritzman, 2004). Goodwin (2000a) also suggested that when useful but difficult-to-model, non-time-series information is available, correcting judgmental forecasts is the most appropriate role of statistical methods. Ahlburg (1984) summarized that the correction procedure is most likely to be useful under three conditions: (a) where a forecaster persistently makes systematic errors, (b) when sufficient history exists for a comparison of actual and forecast values of a series, and (c) where a user does not control forecasts.

2.3.5 Judgmental decomposition

The rationale behind the *decomposition* approach is to split a judgmental task into smaller and less demanding components so that judgmental forecasters can concentrate on one component at a time. Armstrong (1985) suggested that forecasts could be improved by decomposing a subject of interest into a series of more relevant predictions and then mechanically aggregating them to make final predictions.

One main advantage of decomposition is that it can reduce the probability of cognitive biases. However, it should be noted that before a forecaster can take advantage of mechanical rules and decomposition using decision calculus models, managers and/or forecasters should have the cognitive skills to (a) provide a good description of the functional form during the process of generating forecasts and (b) accurately estimate the parameters. Previous research findings in cognitive psychology cast doubt on the ability of forecasters in these areas (Lawrence et al., 2006). Simon (1957, cited in Chakravarti, Mitchell, & Staelin, 1979) found that individuals tended to construct simplified models of situations that made the managers' model look overly simple and incorrectly specified. Other studies have shown that human judgments are frequently subject to systematic bias (Chakravarti, Mitchell, & Staelin, 1979). Thus, even if the forecaster's model is well structured and correctly specified, the parameters required to operationalize the model might be inaccurate if obtained judgmentally (Chakravarti, Mitchell, & Staelin, 1979).

Compared with the previous four integration approaches, relatively less research has been conducted with regard to how decomposition can improve forecast accuracy and the conditions under which it is likely to produce improvements in accuracy (Lawrence et al., 2006). Few studies have investigated the accuracy of holistic forecasts against forecasts resulting from decomposition (Lawrence et al., 2006). For example, in a time-series extrapolation task, Edmundson (1990) found that the combination of separated estimates of the trend, seasonal and random components led to larger accuracy improvements compared to the holistic forecasts. Webby, O'Connor, and Edmundson (2005) also found that people benefit from the use of a decomposition-based decision aid in a task. Similarly, Armstrong and Collopy (1993) found a substantial reduction in forecast error was achieved through structuring the selection of forecasting methods and the weighting of combined forecasts around the judge's knowledge of separate factors that affect the trends in a time series.

In a factorial design, Lyness and Cornelius (1982) tested the influence of information load on the effectiveness of decomposition with the expectation that decomposition would be most advantageous in complex circumstances. The results of their study showed that (a) when assessed by mean absolute deviations, the decomposed judgment strategy was superior overall to the combined forecasting methods and (b) when assessed by correlations, a simple holistic strategy was as effective as the decomposed judgmental approach. However, their study did not fully investigate the impact of decomposition on the relationship between environmental complexity and decision quality. Goodwin and Wright (1993) pointed out that decomposition does not ensure improvements in accuracy and that it actually may harm accuracy under certain situations; for instance, when decomposed judgments are psychologically more complex than holistic judgments, or when increasing the number of judgments required by the decomposition task induces fatigue, or when the decomposition is mechanical and the forecaster is sceptical about the decomposition technique (Goodwin & Wright, 1993).

A number of studies have been devoted to investigating how to improve forecast accuracy in decomposition, but most of these were conducted in laboratory environments. The utility of algorithms as a numerical estimation aid has been evaluated by MacGregor, Lichtenstein, and Slovic (1988). They found that a decision aided by greater structure led to increased forecast accuracy and that subjects performed best when aided with the full algorithmic decomposition strategy. McGregor and Lichtenstein (1991) also used the extended algorithm approach and found that this resulted in an improvement in estimation performance: In their study, subjects trained in algorithmic decomposition were able to produce algorithms, the effectiveness of which was dependent on the presence of misinformation about the components of the quantity to be estimated.

2.4 Judgmental Forecasting in the Tourism and Hospitality Literature[3]

2.4.1 Overview

In the general forecasting literature, forecasts that are based on pure judgment or with judgmental adjustments to the statistical forecasts are commonly known as judgmental forecasts (Wright & Goodwin, 1998). In the tourism literature, terms such as *qualitative*, *intuitive* and *speculative* have been adopted in addition to *judgmental*.

For their forecasts to be of any practical value, tourism planners and decision-makers must adjust their forecasting techniques to deal with a bundle of qualitative factors denoting “expected turning points in a policy framework along a timescale as a result and extension of quantitative data processing” (van Doorn, 1982, p. 161). The judgmental approach is thus designed to incorporate the managerial knowledge

[3] Contents in this section (except for Section 2.4.3) were published in Lin and Song (2012).

of experts into tourism forecasts in order to make more meaningful forecasts that are relevant to managerial decision-making. van Doorn (1982) described judgmental forecasting techniques as being “based on a blend of intuition, expertise, and generally accepted assumptions” (p. 156), which offers the advantage of incorporating expectations about future policy decisions by means of method-implicit procedures.

A deeper understanding of the methodological competence of the different judgmental forecasting techniques will assist tourism analysts and forecasters to make better decisions when choosing an appropriate forecasting tool for a forecasting task. This section summarizes various judgmental forecasting techniques applied in tourism since the 1970s.

Experimentation is often avoided in tourism studies because of the perception of the unnaturalness of the behaviour under analysis (Pizam, 1994). Therefore, this review includes only empirical applications of judgmental forecasting methods in tourism. Depending on the type of participants involved in forecasting techniques, judgmental methods in tourism can be classified into four categories: asking the stakeholders, asking the experts, asking the public, and judgment-aided methods (see Figure 2.3).

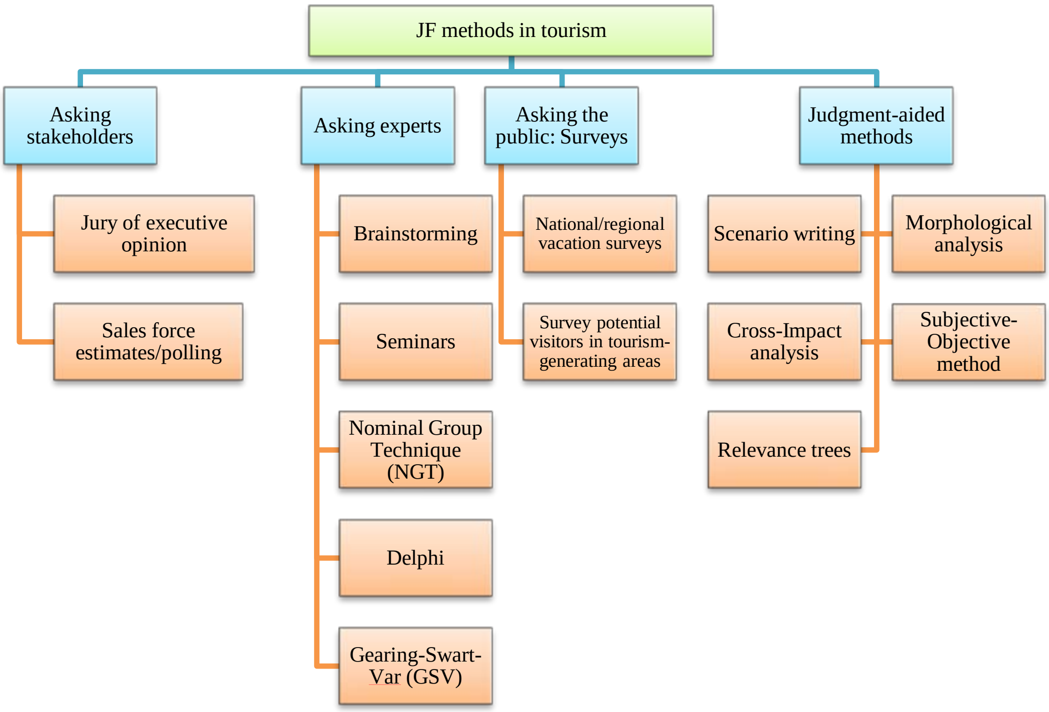


Figure 2.3 Summary of judgmental forecasting methods in tourism

(1) Asking the stakeholders

One of the simplest but most widely used judgmental approaches is the *jury of executive opinion*, which requires little skill or training to participate in and little historical data. It also serves to pool the experience and judgment of those most familiar with the variables to be forecast. Indeed, it is very common for chief executives to seek the opinions of other members of their organizations in order to broaden the base of forecasts and reduce subjectivity. At the micro level, for example, when deciding whether to construct a new restaurant at a particular location, entrepreneurs can sometimes predict demand as accurately as, or even more accurately than, the most rigorous econometric forecasting techniques (Archer, 2000). Given its celerity and simplicity, this method will remain a popular forecasting technique for private enterprises such as individual facilities, attractions, and destinations (Frechtling, 2001). One variant of this method is to obtain group estimates by participants via paper-and-pencil work and then combine them to produce an average estimate. This can be regarded as an informal variant of the Delphi method, the key difference being the lack of a mechanism to prevent interaction among participants. Examples of this executive judgment method include UNWTO's invitation in 1998 to its 211 member countries and territories and 50 international industry practitioners and academic researchers to contribute their views in order to develop forecasts of tourist arrivals between 44 pairs of subregional country groupings up to 2020 (Frechtling, 2001) and the UNWTO panel of tourism experts, where more than 250 experts contribute information on tourism trends (Goeldner & Ritchie, 2005).

One most damaging limitations of this technique, however, is that often the most forceful executive's opinion will dominate the group discussion, which probably reduces the forecasting ability of the whole group. Archer (2000) indicated that unless such discussions are structured, this process could deteriorate into "a guessing game" (p. 63). With this problem in mind, Moutinho and Witt (1995) applied the jury of executive opinion over the Delphi method to rank the importance and probability of the occurrence of 25 possible future developments affecting the world tourism industry and to predict the most likely years of occurrence up to 2030. They argued that it was important to permit a thorough group discussion to facilitate the interchange of ideas and clarifications of reasoning before making forecasts owing to "the radical nature of some of the proposed developments" (Moutinho & Witt, 1995, p. 49).

Another drawback of this method is that it usually provides point estimates of future variables as the most likely forecasts (Frechtling, 2001). It is often easier for a judge to suggest a probability distribution than to give a single future value. This method is called *subjective probability assessment*, but it has very few applications in the tourism field.

Unlike the jury of executive opinion approach, the sales force estimates method does not analyse or amalgamate the predictions of stakeholders (e.g. travel agents or tour operators) from the tourism industry in order to examine their intentions or assess their practical forecasts of future demand (Archer, 1980). Instead, such forecasts benefit from the specialized knowledge and experience of sales representatives and sales managers and may provide meaningful forecasts for the short term, which in turn may help to reinforce self-fulfilling prophecies by means of imposing travel-capacity constraints (Archer, 1980, 2000).

(2) Asking the experts

In a more common approach, a panel of experts is brought together to reach a consensus on a particular event or question. One basic technique, which should be combined with other advanced methods (e.g. scenario writing), is *brainstorming*. The use of this collective inspiration stimulates creative thinking and considers unconventional alternatives that may be unrestrained by present norms and values (Whyte, 1992). Although the brainstorming method considers many alternatives, it might be difficult to apply or relate to the real world. Instead, seminars are frequently used in tourism. For instance, after obtaining forecasts from its member countries and territories in the aforementioned example, UNWTO conducted follow-up regional seminars to present all of the forecasts with the aim of arriving at a consensus on the growth rates of inbound flows expected among the subregions to 2020.

Expert opinion may also prove useful for discovering themes and issues and is often correct about likely results in inexact areas of study; its forecasts can also be accurate (Whyte, 1992). Manning and Fraysier (1989) found that responses on the same questionnaire evaluating recreation issues between state experts on recreation and a representative sample of state residents differed on half of the items. They concluded that experts and other leaders tend to take a more coordinated, institutional view of community services, while the public tends to have a greater exchange or production/consumption orientation, both of which are valid and necessary to achieve viable recreation planning. The choice between using individual versus group techniques thus really comes down to the particular situation. In tourism studies, three such group techniques are commonly used, namely, the nominal group technique (NGT), the Delphi technique, and the Gearing-Swart-Var (GSV) method.

The Delphi technique

Originally developed by the RAND Corporation, the Delphi technique is a valuable working tool for both the long-term planning and the long-term forecasting of tourism development (Cunliffe, 2002; Dalkey & Helmer, 1963). Ng (1984) described it as "the systematic utilization of the judgment of experts [that] aims to obtain consensus among judges on informed predictions of future events" (p. 48). This technique is perhaps the most formalized and studied of the structured group approaches (Wright & Goodwin, 1998). It is also well known for the following: its anonymity of response, iteration, and controlled feedback; convergence in the distribution of opinions as a consequence of the feedback of information; and statistical group response (e.g. median, mean). A detailed review of Delphi forecasting studies in tourism is presented in Section 2.4.2.

Nominal group technique (NGT)

Unlike Delphi, NGT requires the assembly of participants in one location so that members are not anonymous and communication occurs directly between them (Liu, 1988). Developed by Andre Delbecq and Andrew Van de Ven in 1968 (Delbecq, Van de Ven, & Gustafson, 1975), NGT is a special-purpose technique used in behavioural science to tap the ideas and judgments of individuals while simultaneously reaching a group consensus. Ritchie (1994) presented modern applications and forecasting situations where NGT was applied as a useful forecasting tool for tourism analysts. In 1984, Travel Alberta and the Tourism Industry Association of Alberta applied NGT to identify priority issues and problems facing tourism in Alberta. As part of a three-phase study, NGT was designed to determine the views of the private sector concerning provincial tourism development and promotion (Ritchie, 1985). Another practical application of NGT is provided by the US Federal Aviation Administration (FAA), which sponsors an international Forecast Assumptions Workshop that periodically invites some 120–140 industry planners and forecasters representing airlines, aircraft manufacturers, engine manufacturers, trade associations, academic institutions, and other industry groups to critically evaluate the techniques and practices used by the FAA and other aviation forecasters and to examine the outlook for the aviation industry and its growth prospects (FAA, 2004). Workshop participants are divided into several subgroups and are then instructed to critique FAA aviation forecasts for their specific areas. Each subgroup is asked to identify specific assumptions about the short- and long-term future trends of the economic and aviation variables important to their segments of the industry, to indicate why these trends are considered important, and to explain why specific trends are anticipated. At the end of each group discussion, attempts are made to reach a consensus and the most likely future course of these variables.

Gearing-Swart-Var (GSV)

Like Delphi and NGT, the GSV technique also relies on expert opinions. This technique is particularly useful when it is hard and expensive to collect primary data and desirable to use expert judgment as a proxy (Calantone, Benedetto, & Bojanic, 1987). It has also been found to have high validity in real-life applications (Var, 1984). Liu and Auyong (1987, cited in Liu, 1988) offered such a successful application in Turkey, British Columbia, and Hawaii, where they used this method to determine the relative attractiveness of various tourist attractions in each location and to recommend optimum resource allocations in the tourism sector. One major limitation of this technique, however, is that experts are interviewed individually and there is no feedback or consensus (Kaynak & Macaulay, 1984). Furthermore, this technique is too specialized for “general trend and issue determination”; it is also “less powerful than the Delphi technique” (Whyte, 1992, p. 202).

Selecting the most suitable forecasting method using experts in tourism depends on evaluating the level of uncertainty involved, the level of forecast accuracy required, the availability of resources, and the time needed to obtain the forecasts. For example, NGT would probably be the favoured when accuracy is critically important and cost is not a major concern, but it may also be both difficult and expensive to bring groups of experts together for a face-to-face meeting. Alternatively, it may be more appropriate to use a postal Delphi. However, if a trade-off with slight accuracy reduction is allowed, statistical group techniques, which are often much simpler and less costly, can be used. Moreover, when the potential exists for major or discrete changes, scenarios can be incorporated into either the NGT or Delphi process, providing a convenient framework for assessing the potential impact of the subjects being investigated.

(3) Asking the public: Surveys

The judgmental forecasting methods discussed so far rely on experts, whereas an alternative is to ask actual purchasers to provide opinions on their future demand — in other words, to survey consumers as to whether they anticipate taking a trip over the short to medium term (Frechtling, 2001). Two traditional approaches frequently used to seek consumer opinions are the “analysis of national or regional vacation surveys” and “survey inquiries of potential visitors in tourism-generating areas” (Uysal & Crompton, 1985, p. 8). The first type of survey, usually less expensive, may provide valuable information about emerging trends, while the second type may offer useful insights into the attitudes or prevailing images of the potential market toward a tourist-receiving destination (Uysal & Crompton, 1985).

Buying intention surveys, which assume that consumers can predict their purchases in the case of consumer durable goods because they tend to get involved in long-term planning for durable purchases, have been widely used to produce sales forecasts (Huth, Eppright, & Taube, 1994). Lee, Elango, and Schnaars (1997) concluded that the successful usage of buying plans as a valuable forecasting tool had only been found when the analysis was jointly done with economic data. They also empirically tested the efficacy of the Conference Board (CB) survey as a useful forecasting tool. Since 1967, the CB of New York, in its Consumer Confidence Survey of Buying Plans, has asked a question about intentions to take a vacation trip within the next six months (CB, 2011). Since 1977, the CB has interviewed a number of respondents typically 2,500 to 3,500 from among about 5,000 households.

The CB recognized the problem from using such a non-probability sample and started to use a probability-design random sample with post-stratification weights and the US Census X-12 seasonal adjustment from February 2011 (CB, 2011). This type of survey might still serve as a good guide to trends or turning points in future vacation travel activity.

Similar national surveys conducted in Australia since the early 1990s have asked about the leisure activities respondents would have liked to participate in during the survey period but had not been able to. Another example is the quarterly national online survey (or the *Travelhorizons*™) using a sample of nearly 2,300 respondents from a database of over 32,000 US adults and travellers conducted by the US Travel Association (USTA, 2011) since 2007. This survey claims itself to be the first and only tracking survey designed to measure the effects of current economic, social, and natural developments on both the leisure and business travel intentions of US residents over the next six months (USTA, 2011). Questions such as intentions to travel for leisure purposes, reasons for not taking a leisure trip (e.g. time constraints), intentions to travel by census region, and leading leisure travel indicators (e.g. intention to take a leisure trip in the next six months) are included in the survey. Veal (2010) showed that the results from such surveys are useful for market intelligence and in some circumstances could be an indicator of possible future trends in behaviour but could not necessarily be regarded as actual values.

Dwyer, Forsyth, and Dwyer (2010) concluded that forecasts derived from surveys are generally more reliable in the short to medium term than in the longer term. However, they also indicated that the accuracy and reliability of forecasts based on surveys depend on the quality of the survey instruments, the quantity and quality of the responses, and the interpretation of those responses. Drawing on the forecasting performance of two consumer intention surveys, Frechtling (2001) concluded that surveying consumers about their future travel plans may appear to be a reasonable source of valid information about future tourism demand but does not always predict that demand accurately. For example, the vacation intentions model based on the CB's bimonthly survey did not perform well on any of the three accuracy criteria (i.e. error magnitude, directional change, and trend change). The MAPE of the forecasts produced by CB was 2.4 per cent, higher than the forecasts produced by a seasonal Naive model. Lee, Elango, and Schnaars (1997) used the longer time span of 1978 to 1992 to compare the CB's forecasts with those generated from the Naive model and the simple six-month moving average model. They found that the overall MAPE for the intention-to-buy forecasts was nearly double that of the Naive model, and even more than one-third higher than that of the moving average. The comparison showed that the performance of the judgmental forecasts obtained from the CB approach lagged badly relative to that of two very simple extrapolation methods. Frechtling (2001) noted two types of errors that could render intentions invalid as indicators of future tourism-related behaviour: sampling errors (resolved by achieving high response rates) and response errors (resolved by encouraging respondents to answer carefully constructed, practicable questions honestly and objectively). Lee, Elango, and Schnaars (1997) showed that the judgmental approach may be better used to predict "the direction of change rather than the magnitude of the change" and may have less ties to past patterns, whereas extrapolation methods "simply project past trends" (p. 130). Frechtling (2001) also suggested that "a consumer intentions survey that focuses on activities of value to tourism planners and marketers that can be accompanied by a sound time series of actual behaviour will prove a fruitful source of tourism demand forecasts in the future" (p. 233).

(4) Judgment-aided models (JAM)

Two most popular judgment-aided forecasting approaches used in tourism and hospitality contexts are *scenario writing* and *morphological analysis*. Examples of their applications are elaborated below.

− *Scenario writing*

The most quoted definition of a scenario is given by Kahn and Weiner (1967) as "[a] hypothetical sequence of events constructed for the purpose of focusing attention on causal processes and decision points" (p. 6). A scenario is also defined as "an account of what could happen, given the known facts and trends" (Vanhove, 2011, p. 200) or as "a series of events intertwined to form a concept of the future" (Moeller & Shafer, 1994, p. 474). van Doorn (1986) described the scenario technique as follows: "a scenario gives a description of the present situation, of one or more possible and/or desired situation(s) and of one or more sequences(s) of events which can connect the present and future situations" (p. 36). It is evident that a complete scenario under such definition contains at least three central components: a dynamic description and analysis of an existing situation (*baseline analysis*), potential future situations (*future images*), and development lines into the future (*future paths*) (Calantone, Benedetto, & Bojanic, 1987; Dwyer, Forsyth, & Dwyer, 2010; van Doorn, 1986). The underlying assumption of scenario writing is that the "future is not merely some mathematical manipulation of the past, but the confluence of many forces, past, present and future, that can best be understood by simply thinking about the problem" (Schnaars, 1987, p. 106). This approach seeks to generate new information through discussions of an issue by a panel of experts supported by previously produced quantitative evidence. For instance, in demand forecasting, a hypothetical sequence of events is described showing how demand is likely to be affected by a particular causal process. The intent is to indicate what actions can be taken to influence the level of demand at each stage and what the repercussions of such actions might be (Uysal & Crompton, 1985).

Scenario writing usually provides a more qualitative and contextual description of how the present will evolve into the future and identifies a set of possible futures (Schnaars, 1987). Table 2.2, which summarizes a select number of scenario writing studies, shows that scenarios built in tourism research have been used to depict different assumptions or expectations of future growth. The number of scenarios in these studies ranges between two and five, although Schnaars (1987) suggested that the optimal number of scenarios to generate is three. van Doorn (1986, p. 39) summarized four forms of scenarios that have been frequently mentioned "forecasting techniques disguised as scenarios", "parameter variations of one single variable", "variables related to sector developments", and "alternatives for societal developments". One example of the first category is found in a study by Tesar, Edgell, and Seely (1979), who applied a modified scenario research method to develop a slightly optimistic scenario of the impact of Western German tourists on the economy of the USA. The scenario used in their study was not the same as that defined by van Doorn (1986) because it did not contain any single element of the three-component scenario concept.

The second category of scenarios, which takes parameter variation as a scenario, also has little to do with the three-component scenario definition as it provides neither a baseline analysis nor future images. van Doorn (1986) stated that the term "future paths" embodied considerably more than just fluctuations of the parameter value of a tourism variable; rather, it aimed at a very complex system of intended and unintended actions. However, parameter fluctuations can still be used as inputs for tourism scenarios, in particular for elaborating the approximated course of future paths. In addition, scenarios can be incorporated into the results of quantitative models. For example, Smeral, Witt, and Witt (1992) generated multiple forecasts of tourism imports and exports over the period 1991–2000 for a complete system of demand equations using three different scenarios (see Table 2.2). Such procedures are essentially quantitative and mechanistic in nature, but are still taken as “scenario analysis” since more than one forecast is produced (Schnaars, 1987). Similarly, Smeral and Weber (2000) incorporated two scenarios of the EU’s growth path into a model as forecasting assumptions. Smeral and Weber (2000), however, claimed that caution should be exercised in using the forecasts since they captured only the indirect effects of the monetary union, such as those of stable exchange rates and growth. Also, to examine the impact of the global financial and economic crisis on Hong Kong tourism demand from the top 10 source markets over 2009–2012, Song et al. (2010) constructed four scenarios, from the most pessimistic to the most optimistic, according to different assumptions of income levels and tourism prices in those top source markets (see Table 2.2). This technique may provide important stimuli in raising stakeholders’ awareness of different tourism scenarios which might affect policymaking and acceptance (van Doorn, 1986).

The scenarios sketched in the third category are well considered variations of tourism developments based on trend extrapolations of historical tourism developments. A good demonstration of such work is conducted by the Hudson Institute, which applies a two-component scenario model: a baseline analysis component reviewing the development of tourism in the past and a future image component outlining the possibilities for tourism in the future. van Doorn (1986) criticized the use of such a scenario as it takes tourism as a system in itself and ignores the impact of societal developments on tourism.

The last category of scenarios concentrates on alternatives for societal developments in which tourism is considered as one of the subsystems of society. van Doorn (1986) observed that there was a lack of consensus in the use of the three-component scenario definition and suggested using it to judge existing tourism-related scenarios. It is found that not all scenario applications in tourism studies incorporate all three constituent components: at least one or, more often, two are not included. One example is the work of Schwaninger (1989) who constructed scenarios about likely future trends in leisure time and tourism between the years 2000 and 2010. In particular, his study dealt with a base scenario that portrayed the most likely trends by analysing the interactions between economic, political, socio-cultural, ecological, and technological factors. van Doorn (1986) classified his work as a one-dominant-component scenario because Schwaninger (1989) solely emphasized one component of future image, although he also provided information on the component of future paths, but in much less detail. The one-component scenario discussed above deals with only one future image along with a vaguely described future path. BarOn (1975, cited in van Doorn, 1986) extended the use of scenarios by adding alternative future images based on alternative assumptions. His study on forecasting tourism in Thailand, however, still lacked a visible link from the past to the present or the component of future paths. Bearing this in mind, tourism researchers have been more cautious about the construction of scenarios using all of these three components together. Scenarios have been built upon a number of general socioeconomic factors, such as economic growth, income level, and inflation. Koster (1979, cited in van Doorn, 1982), for example, provided forecasts about seven tourism-related fields (i.e. economy, leisure time, population, nature, space, technology and science, and politics) and then correlated these forecasts in a systematic framework to generate a weighted prediction regarding the consequences of the correlations for tourism development in the Netherlands.

Drawing on alternative assumptions regarding the environment for international tourism, BarOn (1984, cited in van Doorn, 1986) produced three scenarios (optimistic, intermediate, and pessimistic) for tourism in Thailand from 1975 to 1980. The fields of interest to these scenarios were focused mainly on political factors, economic tourism development and promotion, and air transportation. More recent studies can be found in Table 2.2. van Doorn (1982), however, pointed out that the scenario writing technique needs to be "supported by more elaborate techniques that will enable the forecaster to improve his assumptions, to strengthen their predictive power, and to widen their scope to range over qualitative data" (p. 163).

Table 2.2 Studies of scenario writing in tourism research

Study	No.	Description of scenarios
Tesar, Edgell, & Seely (1979)	5	Five scenarios (optimistic, slightly optimistic, neutral, slightly pessimistic, and pessimistic) built on three sets of factors influencing tourism: institutional, functional, and product or service.
van Doorn (1986)	4	1. Conventional success: no change in the economic growth and value system. 2. Transformed growth: selective economic growth and transformation of social values. 3. Frustration: stagnated economy and conventional values. 4. Self-restraint: economic decline and transformation of social values related to quality of life spheres.
Martin & Mason (1990, cited in Vanhove, 2011)	4	Four scenarios based on the dimensions of economic growth (high/low) and social attitudes (conventional/transformed) to forecast leisure trends in the UK.

Table 2.2 Studies of scenario writing in tourism research (Continued)

Study	No.	Description of scenarios
Yeoman & Lederer (2005)	4	1. Dynamic Scotland: high disposable income, favourable exchange rates, leading international tourism destination, 7% growth in tourism expenditure, etc. 2. Weekend Getaway: Strong competition for disposable income, favourable exchange rates that attract European visitors, lots of competition from other destinations, attractive leisure destination, 4% growth per year, etc. 3. Yesterday's Destination: unfavourable exchange rates and outbound tourism, decline in international tourism, uncompetitive and expensive destination, substantial decline in the short-break market, second-division destination, 1% growth per year, etc. 4. Exclusive Scotland: no disposable income, favourable exchange rates, weak domestic tourism, international luxury and exclusive resorts, 4% decline per year, etc.
Song et al. (2010)	4	1. Scenario A (most pessimistic): GDP declines 3% over 2009–2010 and grows at 1% over 2011–2012; no change in prices. 2. Scenario B: GDP declines 1% over 2009–2010 and grows at 3% over 2011–2012; no change in prices. 3. Scenario C: GDP declines 3% over 2009–2010 and grows at 1% over 2011–2012; prices decline 1% over 2011–2012. 4. Scenario D (most optimistic): GDP declines 1% over 2009–2010 and grows at 3% over 2011–2012; prices decline 1% over 2011–2012.
Varum, Melo, Alvarenga, & de Carvalho (2011)	4	Depending on the four uncertainties of client dynamics and loyalty, territorial planning and sectoral regulation, industrial structure, and Portugal's attractiveness as a tourist attraction, four scenarios are built: Portugal – southern experience; Portugal – global emotions; Portugal – "sin surprise"; and non-charming Portugal.
Smeral, Witt, & Witt (1992)	3	1. Baseline scenario: no change in the external environment. 2. European Community (EC) completion scenario: completion of a single internal market of the EC taking place at the end of 1992. 3. Growth scenario: increased world growth likely to result from the liberalization of Eastern Europe.
Rossetto (1999)	3	1. Rapid return to growth (baseline forecasts): assuming Japan gradually opens its economy, while Thailand, Indonesia, Malaysia, and South Korea experience a period of contraction and stabilization. 2. Steady return to growth (most optimistic). 3. Slow return to growth (most pessimistic).
Patterson & McDonald (2004)	3	1. Scenario A: no technical change over the period 1997–2007. 2. Scenario B: mid-range technical change over 1997–2007 based on the exception of some slowdown in historical rates of technical change. 3. Scenario C: continuation of historical levels of technical change over 1997–2007.
Smeral & Weber (2000)	2	1. The "business-as-usual" case: assuming limited progress in trade and investment liberalization. 2. The "high-performance" case: assuming more progress and a higher pace of structural reform.
Yeoman, Lennon, & Black (2005)	2	Two scenarios representing the stages, events, and communications that would occur in the event of a suspected outbreak and a confirmed outbreak of foot-and-mouth disease in Scotland.
Tolley, Lumsdon, & Bickerstaff (2010)	2	1. A "business-as-usual" scenario (cheap, private motorized mobility; reliance on techno-efficiency, etc.). 2. Sharp increase in the price of motorized transportation (peak oil, carbon taxes, generalized road pricing, etc.).
FAA (2010, 2011)	2	1. Optimistic forecast: lower inflation and faster growth in the labour force and capital stock than in the baseline forecast. 2. Pessimistic forecast: higher inflation and slower growth in the labour force and capital stock than in the baseline forecast.

Scenario writing methods can be qualitative or quantitative as well as some mixture of both. The strongest proponent of the qualitative approach to scenarios has been Kahn (1979), who developed scenarios for the future of the USA and the world based on narratives and who predicted that in 2000, tourism would be the largest industry and the most important export sector in the world (Witt & Moutinho, 1989). van Doorn (1986) used quadrants and two languages (English and French) to describe the typology of scenarios and established their relationships: (a) projective and prospective scenarios, (b) normative and descriptive scenarios, (c) dominant and "limits-identifying" scenarios, and (d) preferential and aprioristic scenarios. He clarified that prospective scenarios are always normative whereas projective scenarios could either be descriptive or normative. van Doorn (1982) further showed there are no great differences between the projective and prospective scenario writing. The desired state described in the normative scenario might cause considerable difficulties, and even if such problems can be solved through an acceptable solution, it is likely to still have certain methodological problems (e.g. treatment of consistency, plausibility, and level of aggregation as a challenge). van Doorn (1982) also observed that there were only a few noteworthy applications of exploratory scenario writing in the tourism field owing to "the novelty of the technique (relatively speaking) and the difficulty in handling qualitative data with tools developed for quantitative data processing" (p. 161).

Scenario writing is not a real forecasting technique in itself but can be used to develop medium- to long-term scenarios whose likely eventualities can then be analysed for their potential effects upon tourism demand (van Doorn, 1984). Thus, it can create valuable input for group forecasting, such as with the Delphi technique. It may also be applied to a future determined by the Delphi approach, examples being

Henderson and Bialeschki's (1984) study of organized camping and the future and Tolley, Lumsdon, and Bickerstaff's (2010) study of future walking.

Scenario writing is particularly useful for examining the likely impact of changes of greater magnitude, such as crises or large-scale policy changes (Dwyer, Forsyth, & Dwyer, 2010). On the basis of a scale of severity, probability, type of event, and level of certainty, Prideaux, Laws, and Faulkner (2003) developed a framework to classify group shocks into four categories (S4: "not anticipated"; S3: "unlikely but just possible"; S2: "the possible based on a worst-case scenario of past trading conditions"; and S1: "the expected based on recent past trading conditions"). It was recommended that scenarios be applied under the assumptions of S3 and S4. Yeoman, Lennon, and Black (2005) examined how a future outbreak of foot-and-mouth disease in Scotland would be treated and considered the potential reaction by government agencies with particular reference to communication and the management of crises within the tourism sector. Delphi forecasting may thus be useful in developing estimates of post-shock travel demand and supply conditions. The value of using scenarios is that by considering potential developments and responses in advance, an organization will not be forced to make quick, ill-considered decisions when such unexpected conditions occur (Faulkner & Valerio, 1995). In some other cases, forecasts produced by scenario writing have been so vague or trivial that one might wonder whether they could benefit future planning. One such example is Kahn's (1979) study, which predicted that in 1989, the tourism growth rate would be double the economic growth rate; this result was not surprising, since it had already been recognized in the literature (van Doorn, 1982).

The use of scenario projections is not only attractive to academics but has also been widely applied in real-world forecasting by tourism organizations and industry stakeholders to construct powerful policy visions. Adopting a scenario approach, the Tourism Forecasting Council postulated three scenarios (rapid/steady/slow return to growth) representing possible future conditions in the global economy to predict future tourist arrivals (Rossetto, 1999). To describe the environmental implications of national tourism forecasts in New Zealand, Patterson and McDonald (2004) developed scenarios to construct projections of future resource use and pollution by the tourism sector from the base year of 1997 to 2007. Their study produced three scenarios highlighting the difference between three levels of technological improvement. Patterson and McDonald (2004) also explained their reasons for using scenarios rather than forecasts, namely their consideration of unpredictable events, which would have made any forecasts highly prone to errors, and their inclusion of environmental (resources and pollutants) variables, which would have made predictive forecasting very difficult and problematic. Lennon and Yeoman (2007) examined how the National Tourism Organization of Scotland (VisitScotland) utilized the scenario planning approach to capture expert opinions. “What if” thinking was applied to paint the future, and conclusions were drawn from two potential future scenarios. They found that the future of Scottish tourism to 2015 would be affected by macrotrends and drivers (e.g. globalization, sustainability, technology/communication, politics, etc.) in UK society. To reflect uncertainties in projecting economic growth, the FAA Aerospace forecasts built three scenarios to produce base forecasts of aviation demand and activity levels as well as high and low economic growth cases (FAA, 2010, 2011). The base forecasts were generated from econometric models, while the high and low economic growth rates were based on optimistic and pessimistic scenarios from Global Insight’s *30-Year Focus*.

To conclude, the basic purpose of scenario writing is to provide multiple forecasts; therefore, it makes more sense to establish a number of plausible assumptions rather than rely on a single one that may later turn out to be incorrect (Schnaars, 1987). van Doorn (1986) stressed the need to seek consensus on a common scenario methodology. Although scenario writing has been adopted primarily for medium- and long-term forecasting, there is no empirical evidence to indicate that it would not be suitable for shorter term forecasts (Schnaars, 1987). To date, very few tourism studies have examined the relative accuracy of scenario forecasts or made direct or indirect comparisons with other judgment or quantitative methods. Combining scenario analysis with other forecasting techniques, such as time-series analysis, Delphi, and cross-impact analysis, is recommended by van Doorn (1986).

– *Other judgmental forecasting techniques*

Alternative judgmental approaches, such as morphological analysis, cross-impact analysis, relevance tree analysis, and the subjective-objective method, have also been proposed and used in tourism forecasting. The first three methods share a similar way of presenting their outputs but with different structures and content, in the form of a matrix. The matrix produced from the morphological analysis (or a *morphological box*) explores all possible solutions to a multidimensional, nonquantified, and complex problem (Uysal & Crompton, 1985), the results of which are qualitative in nature. Management looks closely at potential combinations and assesses the various attainable levels of demand under different assumptions about the performance of each variable (Uysal & Crompton, 1985). Although it has been argued that this technique lacks rigour unless supported by numerical analysis, it can provide valuable input for group forecasting discussions (Archer, 2000).

An extension of the Delphi technique, cross-impact analysis involves identifying and evaluating the impact of trends or events upon one another using a matrix format, thereby enabling tourism managers to gain deeper insight into the sensitivities and interrelationships among a number of policy options (Archer, 2000). Data are collected by asking participants to attach probabilities to events occurring in the future and to consider how each probability is affected by each event (Archer, 2000). In much the same way as the Delphi technique, this method also depends on the ability of experts to provide meaningful estimates of the probability an event will occur. Some strengths of such a technique are that it (a) provides "form and structure to quantitative predictive models", (b) integrates the "interests of a wide array of public" through distilling "conventional wisdom and collective judgment" in order to arrive at a consensus, and (c) can handle "complex issues where no clear consensus or interaction is available" (Whyte, 1992, pp. 201-202). Similar to Delphi, this technique also suffers from the problem of direct expert influence. Becker et al. (1985, cited in Whyte, 1992) provided one example by applying this technique to identifying possible trends and events affecting the southeast region of the US National Parks Service in the 1990s. Additionally, this technique may be time-consuming if several iterations are required; also, if the matrix is very large, it may not reflect reality and so yield insufficiently consistent responses. Unlike these two methods, the output matrix in the relevance tree approach is in the form of a visually hierarchical structure exhausting all possible ways of achieving objectives. Although these three methods have been used in other areas of forecasting, they have not yet been widely applied in tourism.

The subjective-objective method was initially proposed and developed by Ng and Knott (1979, cited in Ng, 1984) to ascertain future manpower needs for leisure services. Ng (1984) presented a forecasting framework consisting of three components (multiple regression submodels, the subjective–objective forecasting model, and the Delphi method), to forecast the demand for leisure services manpower. From a methodological perspective, this method does not rely on historical relationships among variables of interest or on the assumptions that these relationships will continue into the future. Instead, it distils both the practical and the professional knowledge of each chief administrator regarding an organization's or company's unique situation and future plans. However, as revealed by Ng (1984), one serious limitation of this technique is that "it is quite impossible for an individual administrator to estimate correctly the effects of changing society needs on manpower situations" (p. 48); also, its outlook on the future is likely to be biased toward the optimistic or pessimistic extremes. Despite these limitations, Ng (1984) suggested that this method could be supplemented by other forecasting approaches as a short-term forecasting tool.

2.4.2 The Delphi technique

The Delphi technique is well established as a judgmental forecasting tool for tourism studies, but it is also the one that has been subject to the most criticism. In a Delphi study, experts are selected from different parts of the tourism and hospitality industry, such as industry operators, public policy makers, tourism and travel associations/organizations, and government tourism departments, and from the general public. It can provide information regarding the future that other conventional extrapolative techniques cannot reliably forecast. Smith (1995) considered the Delphi method to be "one of the best known and sometimes more controversial forecasting methods for tourism futures research" (p. 145), while Kaynak and Cavlek (2007) called it "the cornerstone of futures research" (p. 111).

The Delphi approach is helpful where data are insufficient, changes in a previous trend are expected, or new elements might interfere, with the result that mathematical-type analysis might be inappropriate. Usually, the aim is to provide an indication of the likelihood of specific future events or trends occurring and the probability that these events will occur during a specific time period, most commonly within a 5 or 10 year period (Ng, 1984). Table 2.3 summarizes the main characteristics of the studies published on the Delphi approach since the 1970s. Clearly, this technique has tremendous potential use in both qualitative and quantitative tourism research.

(1) Task(s)/purpose(s)

The Delphi technique has been applied in a variety of locations, the most popular researched region being Europe (14), followed by the USA (10), Asia (e.g. China, Cambodia, Hong Kong, Japan, Singapore, and South Korea) (11), the UK (7), Australia (3), and Canada (3). In the 48 post-1970 empirical studies reviewed in this study, the Delphi forecasting technique was mainly applied in four areas: (a) *event forecasting*, (b) *forecasting tourism demand variables*, (c) *forecasting future trends/market conditions* (the most popular application), and (d) *issue identification/prioritization*.

Table 2.3 Summary of Delphi forecasting applications in tourism and hospitality studies

ID	Study	Region	Panel components	Task(s)/Purpose(s)	R	Panel size	Convergence	Feedback & analysis	Media
1	Dyck & Emery (1970)	Alberta, Canada	Six panels from different areas: social goals and values, the needs of the individual, political life, family life and child rearing, leisure and recreation, and intercultural relations.	Predict future distribution of work and leisure time and the most likely uses of this leisure time in Alberta, Canada over the years 1975 to 2005; forecast the probable dates and probabilities for the occurrence of events associated with leisure and recreation; and project trends using three 10-year time periods and their probabilities.	3	305,149, 126	Modal position (the average for the total panel)	Median dates, trend forecasts in graphical form, reasons and arguments presented together in scenarios.	Phone, by person
1	Shafer, Moeller, & Getty (1974), Shafer & Moeller (1974), Moeller & Shafer (1983, 1994)	USA	Experts from public land-management agencies, educational institutions, communications industries, public regulatory agencies, legislative bodies, quasi-public environmental organizations, and industry.	Probe for social, managerial and technological events that are likely to shape the future of park and recreation management to the year 2000.	4	405 in r4	Median & interquartiles	Individual estimates, median, interquartiles, graphic summary of the dates range, reasons (of those with responses outside quartiles)	
1	English & Kearnon (1976)	USA	Experts from major airlines, aerospace industries, national and international regulatory agencies, government & private agencies, and aviation publications.	Predict the air traffic and the developments in aircraft technology.	2	28,23	Coefficient of variation, SD	Mode, mean, lower and upper bounds, SD	
1	Hawkins, Shafer, & Rovelstad (1980), Seely et al. (1980), Kibedi (1981)	World	Participants of an international tourism symposium in Washington, DC in 1979.	Forecast the likelihood of 14 tourism-related events by 2000, the year of probable occurrence, and the impact of events on tourism.	2	25,19	SD	Mean, SD, ranking order for each event statement	Mail
2	Edgell, Seely, & Iglarsh (1980)	USA	A group of tourism experts.	Adjust time-series forecasts of the number of tourist arrivals to the USA and the level of international tourist receipts.			-		
2	Liu (1988)	Hawaii, Oahu	Local experts (tourist receivers) and overseas travel agents (tourist senders)	Forecast the likelihood of possible scenarios, future growth and development of tourism by 2000 (e.g. visitor arrivals, the share of domestic arrivals and Oahu's share of visitors, the visitor-to-resident ratio, and visitor accommodation supply)	2	23,17	Wilcoxon rank sum tests	Individual estimates (subjective probability), mean, SD, standard error, mode, range, quartiles, No. of responses and non-responses, reasons	Mail, phone

Note: R denotes the number of rounds in a Delphi survey.

Table 2.3 Summary of Delphi forecasting applications in tourism and hospitality studies (Continued)

ID	Study	Region	Panel components	Task(s)/Purpose(s)	R	Panel size	Convergence	Feedback & analysis	Media
2	Lee & Kim (1998)	South Korea	Experts from tourism- and World Cup-related fields: tourism department in government, tourism academy, travel agencies.	Predict the number of World Cup related and non-World Cup related foreign visitors and total foreign visitors during the 2002 World Cup in Korea.	2	41,41	SD	Mean, median, mode, SD, range, reasons (of those with responses outside of mean/median).	Mail, fax, phone
2	Tideswell, Mules, & Faulkner (2001)	South Australia	Experts from airline management, inbound tour businesses, and hospitality sector management and tourism marketers and planners from the South Australian Tourism Council.	Predict the future tourism industry potential of South Australia to 2005.	2	26,18	Workshop	Mean, mode, group discussion in a full-day workshop	Mail
2	Lee, Song, & Mjelde (2008) (2008)	South Korea	Experts from tourism academia, Korean National Tourism Organization, tourism research institute, and event managers.	Predict the number of domestic and international tourists to an international Expo in Korea in 2012.	2	27,27	Mean, median, SD, skewness, kurtosis	Mean, median, mode, range, confidence intervals, SD, skewness, kurtosis, reasons	
2	Song, Witt, & Lin (2010)	Hong Kong	Managers/directors in tourism and hospitality industry of Hong Kong, academic researchers	Forecast visitor arrivals and the demand for hotel rooms in Hong Kong up to 2012 by considering the influence of the current economic/financial crisis and swine influenza.	2	9,6	SD	Mean, reasons on r1	
2	Lin & Song (2011)	Hong Kong	Managers/directors in tourism and hospitality industry of Hong Kong, academic researchers	Forecast tourist arrivals from six source markets of Hong Kong from 2011-2015.	2	17,16	SD	Median, mean, minimum, maximum, reasons on r1	
2	Song, Gao, & Lin (2013)	Hong Kong	Experts from academic institutions, the accommodation sector (e.g. hotels, resorts), tourist attractions/tourist facilities, travel trades (e.g. tour operators, travel agents), and government offices.	Forecast visitor arrivals from six source markets of Hong Kong from 2011Q2-2015Q4.	2	18,17	SD	Median, mean, minimum, maximum, reasons on r1	HKTDFS
2	Lin (2013)	Hong Kong	Experts from academic institutions, tourist attractions, and government offices.	Forecast visitor arrivals from China to Hong Kong from 2008Q1-2015Q4.	2	11,9	SD	Mean	Email
3	D'More (1976)	Canada	Each of the four panels had experts from different disciplines: government, different sectors of the tourism industry, and universities.	Identify emerging trends, opportunities, and constraints for Canadian tourism in 1986. All relevant aspects were considered, including demand, the availability of resources, environment, social trends, economic trends, technology, and government policy.	2		-	Individual estimates, comments	

Table 2.3 Summary of Delphi forecasting applications in tourism and hospitality studies (Continued)

ID	Study	Region	Panel components	Task(s)/Purpose(s)	R	Panel size	Convergence	Feedback & analysis	Media
3	Kaynak & Macaulay (1984)	Nova Scotia, Canada	Public policy makers, industry operators, trade associations, local businessmen	Predict possible effects of changes in social values on the development of tourism, the changing structure of the tourism industry over 1982-2000, and the future development of the tourism industry in Nova Scotia and assess the impact of any change on the industry itself.	2	111, 44	Information feedback	Mean, mode, SD, direction of change	Mail
3	Yeong et al. (1989)	Singapore	(1) Stakeholders from six sectors of the local tourism industry;(2) Executives who took programs at the National University of Singapore, representing 50% of the local business community, and their counterparts from 12 foreign countries	Forecast the likelihood of 26 tourism-related scenarios on a 0-100% scale, the year of probable occurrence, and the importance of events to tourism development in Singapore on a 1-5 scale.	3, 2	23,19,17; 45,34	-	Individual estimates, mode, reasons	Mail
3	Green et al. (1990a, 1990b); Green & Hunter (1992)	Bradford, UK	Planners, tourism officers, economic development unit personnel, local residents and traders	Identify the likely environmental impact of tourism projects in both rural and urban environments.	2	31,21	SD, coefficient of variation	Individual responses in r1, sample size, range of magnitude attributed to each impact, median, mean, SD, ranking	Mail
3	Kaynak, Bloom, & Leibold (1994)	South Africa	Tourism analysts from eight categories: policy makers, transport accommodation, attractions, travel organizers, industrial and commercial sectors, educators, and industry operators.	Project the future tourism scenario in South Africa to 2000: value changes in society, changing structure of the tourism industry, events having potential impact on tourism and training.	2	50,37		Mean, median, mode, SD, reasons	Mail
3	McCleary & Whitney (1994)	Six Eastern European countries	(1) Travel professionals and (2) tourism educators from the USA and Canada.	Make projections about travel to Eastern Europe and explore consumer motivations and perceived risks to each of the six countries (9-point scale).	3	22,13,10; 24,19,17	Mean	Mean, rankings, own responses in r1	Mail
3	Taylor & Judd (1994)	New Orleans and St. Louis, USA	Administrators/staffs at Southern Queen, individuals from non-competing organizations, and other channel members and industry experts who were willing to participate.	Project major environmental trend categories for South Queen.		10-15	-	Median, interquartile range, reasons for those whose responses fell outside of the interquartile range.	

Table 2.3 Summary of Delphi forecasting applications in tourism and hospitality studies (Continued)

ID	Study	Region	Panel components	Task(s)/Purpose(s)	R	Panel size	Convergence	Feedback & analysis	Media
3	Pan et al. (1995)	Gozo	A group of Maltese and Goztian experts from public/private sector or other relevant tourism-related fields.	Identify potential impact of tourism in Gozo.	2	21,12	Magnitude of opinion changes over rounds.	General analysis of individual responses, comments from r1	Personal interview in r1, mail in r2
		China	Experts from multinational hotel groups.	Predict tourism potential (e.g. growth rate of international tourists flows to China in the next 5 years); assess the tourism competition and the business and economic environment in China.	1	10	Magnitude of opinion changes over rounds.	Mean	
		Belize, UK	Experts were selected on the basis of their knowledge of Belize and their expertise in marketing.	Predict events likely to affect the strategic marketing of Belizean tourism products (e.g. the likelihood of occurrence of events by 2005 on a 0-100 scale).	3	25,14,14	Magnitude of opinion changes over rounds.	Mean, reasons on r2 & r3.	Mail, phone & personal interview.
3	Müller (1998)	Germany, Austria, & Switzerland	Experts were selected on the basis of four quotas: national quota (50 per country), sector quota per country, gender quota per country (at least 10 women), and age quota per country (at least five under 30, at least 25 under 50).	Predict possible future development patterns in long-distance travel and its impact on domestic tourism by 2005.	3	144 in r3	-	Median, reasons on r2	-
3	Obermair (1998)	World	223 selected international experts from 64 countries.	Project long-term global tourism trends for the next 5-15 years.	3				
3	McCubbrey (1999)	USA	Travel agency owners or employees, internet travel agency owners or employees, officers or employees of air travel industry associations, and consultants to the industry	Predict the impact of electronic commerce technologies and disintermediation and reinter mediation on traditional travel agents in the US air travel distribution industry.	3	17		Individual estimates (percentage of market shares),ranking, mean	
3	Lloyd, La Lopa, & Braunlich (2000)	Hong Kong	Managers who (a) had worked in the hotel industry for 15 years or more, held the job of general manager for 5 years or more, and lived in Hong Kong for 5 years or more; (b) planned to continue to live in Hong Kong after 1997; and (c) were members of the Hong Kong Hotel Association.	Predict changes in the Hong Kong hotel industry as a result of the handover from the UK to China in 1997.	3	14	Coefficient of variation	Individual estimates, rankings, mean, SD, and coefficient of variation	Mail

Table 2.3 Summary of Delphi forecasting applications in tourism and hospitality studies (Continued)

ID	Study	Region	Panel components	Task(s)/Purpose(s)	R	Panel size	Convergence	Feedback & analysis	Media
3	Tolley, Lumsdon, & Bickerstaff (2001)	Europe	Experts from research, practice, policy, advocacy, and planning and others with professional interest in five aspects of walking: everyday/utilitarian/transport, leisure/recreation, health/exercise, tourism and other.	Predict walking trends on a 1-5 scale in terms of planning, policy, strategy, image, status, attitudes and behaviour in Europe by 2010.	2	112,72	Mean, median & SD	Mean, median, mode, SD, reasons	Email, fax
3	Weber & Ladkin (2003)	Australia, UK	Experts from destination marketing companies, professional conference organizing companies, conference/convention bureaus, industry associations, conference venues, and academia.	Predict future trends, key issues and competitive forces (all on a 1-5 scale ranging from 'strong disagreement' to 'strong agreement') that would affect the conventions and meetings industry in Australia/the UK over 2001-2005.	3	14,13,11; 12,8,7	Level of agreement	Mean, rankings, reasons	Email, phone
3	Sadi & Henderson (2005)	Saudi Arabia	Two panels of experts from government ministries, food services, hotels, airline industries, travel agencies, and tourist attractions.	Forecast aspects of the future of the Saudi Arabian tourism industry and assess their impact after 2005.	3	20	-	Rankings, probability of the occurrence of each event, reasons	Mail
3	Weber & Ladkin (2005)	UK (1), Australia (2)	Executives from convention centres, convention and visitors bureaus, convention hotels, professional meeting organizers, industry associations, tourist organizations, and destination marketing companies; and academics.	Predict future trends in business, technology, social, and political areas that would impact the convention and meeting industry in the UK/Australia over 2001-2005.	3	12,8,7; 14,13,11	Mean	Mean	
3	Kaynak & Marandu (2006)	Botswana	Experts were selected from the Department of Tourism and the Hotel and Tourism Association of Botswana.	Forecast tourism market potential of Botswana by 2020: (1) expected value changes in society, (2) the changing structure of the tourism industry, (3) probability of occurrence, time of occurrence, and impact of certain events that affect tourism.	2	104,68	SD	Individual estimates on r1, ranking, mean	Mail, phone
3	Kaynak & Pathak (2006)	Fiji Islands	Knowledgeable international and national tourism analysts.	Forecast tourism market potential of Fiji Islands from 2001-2020 (1) value changes in society, (2) changing structure of the tourism industry, (3) probability of occurrence, time of occurrence, and impact of certain events that affect tourism development and training.	2	60		Mean, median, mode, SD, ranking, reasons	

Table 2.3 Summary of Delphi forecasting applications in tourism and hospitality studies (Continued)

ID	Study	Region	Panel components	Task(s)/Purpose(s)	R	Panel size	Convergence	Feedback & analysis	Media
3	Kaynak & Cavlek (2007)	Croatia, Germany	Tourism and hospitality educators, travel agency and travel organizers, members of tourist boards, public sector agencies	Project the future tourism scenario (on a 1-5 scale) in Croatia from 2001-2020: value changes in the society, changing structure of the tourism industry, and events having a potential impact on tourism and training.	1	49 in r1		Mean, median, mode, SD, reasons	Mail, phone
3	Austin, Leeb, & Getzb (2008)	World	Practitioners and educators in inclusive and special recreation.	Identify trends (i.e. programme trends, approaches to programmes and services, financial trends, and professional trends) in special and inclusive recreation (on a 1-5 scale of importance).	4	25,24,25, 24	Agreement of at least 80% of the members.	Mean, rankings	Internet-based, email
3	Katsura & Sheldon (2008)	Japan	Experts from public and private tourism authorities and academic researchers in the field of tourism informatics.	Predict possible uses of mobile tourism applications in Japan by 2015 and investigate the trends of the future: probability of event occurrence (0-100%), year of occurrence, and importance of scenario (5-point scale)	3	23,21,20	Median, pair-sample t test	Median, reasons on r1 & r2.	Phone, email
4	Garrod & Fyall (2000)	UK	Owners and managers of historic properties, officers of heritage-based organizations, consultants, and academics	Investigate heritage managers' perceptions of sustainability issues: missions of heritage attractions, factors influencing pricing strategy, and heritage conservation for funding organizations.	3	17,15,14	Mean, Spearman's rank correlation coefficient	Individual estimates, mean, ranking, direction of change, rank correlation	
4	Vaugeois et al. (2005)	World	Experts from Canadian Association of Leisure Studies, the Travel, Tourism and Research Association; presenters at academic and practitioner conferences; managers of leisure organizations; and researchers from academic institutions.	Identify future research priorities and develop an agenda for knowledge exchange improvements in the leisure and tourism field.	3	84,64,49		Rankings	Email
4	Spenceley (2008)	South Africa	Experts from government, academia, non-governmental organizations, the private sector, and consultancies	Identify factors perceived as essential for sustainable nature-based tourism operating in transfrontier conservation areas (TFCAs) in southern Africa.	3	42,184, 104	Chi-square analysis	Individual estimates, tally, mode	
4	Donohoe (2011a, 2011b)	World	Two groups: ecotourism professionals from government, private industry, and nongovernmental organizations and academics engaged in ecotourism research and education.	Assess the importance of culture for ecotourism, develop a definition for culturally sensitive ecotourism, and identify the barriers and opportunities associated with its implementation.	3	94,79,61	Increase in agreement percentages, mean, SD	Frequency of individual responses, ranking importance	Internet-based, email

Note: 1: Event forecasting (9), 2: Forecast tourism demand variables (8), 3: Forecast future trends/market conditions (25), 4: Issue identification/prioritization (5).

The majority of the published studies on forecasting future trends or market conditions focus on projecting future trend/patterns, identifying opportunities and constraints, and evaluating the potential impacts of some events or future changes, such as value changes in society and the changing structure of the tourism industry, on tourism. Studies on event forecasting aim to solicit expert opinions in order to predict the likelihood and/or time of the occurrence of specific events and their impacts on tourism. These studies often use the Likert scale to design event statements; for example, Hawkins, Shafer, and Rovelstad (1980) used a scale ranging from 0 ("*Never*") to 10 ("*100 per cent likelihood of occurrence*") to rate event statements, and a scale ranging from 0 ("*Not at all important*") to 10 ("*Critically important*") to rank the importance of the events. Table 2.3 shows that this type of studies became less popular after the 1990s and that now event forecasting more often appears as one component task in forecasting tourism market potential; for instance, Kaynak and Marandu (2006) and Kaynak and Pathak (2006) applied the Delphi technique to forecast tourism market potential in Botswana and the Fiji Islands respectively.

Table 2.3 also shows that Delphi is not only used for qualitative forecasting purposes but is also applied to develop forecasts of tourism demand variables. One very early application was offered by Edgell, Seely, and Iglarsh (1980), who invited panellists to revise the forecasts of tourist arrivals and tourist receipts to the USA obtained from a regression model. Liu (1988) utilized the Delphi technique in the context of Hawaii to forecast visitor arrivals, share of domestic arrivals and Oahu's share of visitors, the visitor-resident ratio, and visitor accommodation supply. Song, Witt, and Lin (2010) and Lin and Song (2011) reported two recent studies that applied the Delphi approach to forecasting tourism demand in Hong Kong by the year 2015.

(2) Selection of participants

The application of the Delphi technique allows the use of different levels of expertise − the technique not only makes full use of the answers from top experts but also those from experts in the upper half of the range. At the start of the Delphi process, one of the most important steps is the selection of the panellists. A Delphi panel should consist of individuals who are willing to participate and who have expertise concerning the issue at hand. Wheeller, Hart, and Whysall (1990) recommended the use of balanced panels of experts from different backgrounds. In a similar vein, Rowe and Wright (2001) suggested that a group of heterogeneous experts whose combined knowledge and expertise reflect the full scope of the research issue is preferable to a group of experts focused in a single domain. Kollwitz (2011) pointed out that one problem associated with the Delphi technique is the identification of an appropriate panel of experts who represent the desired balance of opinions, philosophies, and experience. Tichy (2004) argued that foresight exercises should not only base their panels on the “top specialists of the respective field” but also on “a fair mixture of experts of different grades, with different types of knowledge and affiliation” (p. 341). Donohoe (2011a) summarized that the size, characteristics, and composition of a panel should be “governed by the purpose of the investigation” (p. 30).

Table 2.3 shows that in Delphi applications in tourism, the panel often includes stakeholders from different sectors of the tourism industry reflecting a diversity of experience, knowledge, skills, and perspectives: industry practitioners, tourism and hospitality educators (or academic researchers), professionals from tourism industry associations, government ministries, and nongovernmental organizations. Lapage (1994) argued that equipment dealers, travel agents, park managers, airline stewardesses, and almost any group that has constant contact with the travelling public are potential sources of information on tourists' unmet needs and future market conditions/developments and thus should be included in a Delphi panel. Local residents in the researched region could also be included as panellists (see, for example, Green, Hunter, & Moore, 1990a, 1990b; Green & Hunter, 1992). In addition to canvassing expert opinions, the necessity of involving the local community in the decision-making process has been emphasized. Some studies have already acknowledged the role of experts and local residents when using the Delphi method to assess the environmental impact of tourism projects in both rural and urban environments (Green, Hunter, & Moore, 1990a, 1990b). Green and Hunter (1992) incorporated local public opinion through the Delphi technique to evaluate redevelopment at a site in northern Britain.

In addition to a balanced panel, criteria can be set regarding the types of skills required from stakeholders in relation to the specific objectives of Delphi applications (Spenceley, 2008); for example, Lloyd, La Lopa, and Braunlich (2000) selected panellists according to the following criteria: managers who (a) had worked in the hotel industry for 15 years or more, (b) had held the position of general manager for 5 years or more, (c) had lived in Hong Kong for 5 years or more, (d) planned to continue to live in Hong Kong after 1997, and (e) were members of the Hong Kong Hotel Association. In Donohoe's (2011b) study, a panel of 86 professionals, 32 academics, and 39 professional and academic experts was selected on the basis that they satisfied several predetermined "ecotourism expert" selection criteria (e.g. a minimum of 5 years' working experience in the public, government or private sector related to ecotourism and/or nature-based activities; a minimum of 5 years' teaching experience on courses dedicated to tourism/ecotourism, etc.).

There is no consensus on the knowledge or expertise required for a person to be a Delphi panel member (Yeong, Keng, & Leng, 1989). With regard to expertise requirements, Martino (1983) concluded that panel members' expertise and knowledge on the subject matter was the most important criterion for a Delphi study. Some studies have tried to use self-rated expertise to select experts. Kaynak and Pathak (2006), however, found that in terms of analysing Delphi results, information obtained on self-rated expertise is of limited use because no correlation can be found between the rating of a panellist and the deviation of his/her estimate from the mean.

In most of the Delphi applications presented in Table 2.3, panellists were selected through non-probability sampling methods, purposive sampling being the most popular one. For example, in Hawkins, Shafer, and Rovelstad's (1980) study, Delphi participants were selected based on the consideration of international representation. Dyck and Emery (1970) adopted a modified version of a snowballing technique in selecting panel members. In their study, 14 resource individuals were first selected and asked to participate in the survey; in the meantime, each individual was asked to provide a list of 10 or more persons he or she (a) was acquainted with and (b) considered "knowledgeable" and "informed" about leisure and recreation (Dyck & Emery, 1970). From the complete list of prospective experts, the researchers judgmentally selected a panel of 43 participants. Müller (1998) used the quota sampling method to choose qualified participants: national quota (50 per country), sector quota per country, gender quota per country (at least 10 women), and age quota per country (at least five under 30 and at least 25 under 50). To conclude, tourism researchers have recognized the importance of selecting experts from diverse backgrounds but very few of them have attempted to investigate the influence of panel composition and size on Delphi results.

(3) Panel size

Table 2.3 shows that in Delphi applications in tourism since the 1970s, panel size has ranged anywhere from six to over 400. Although panel size is likely to have an impact on the effectiveness of the technique, there appear to be no firm rules governing the number of panel members. The answer to the question of what constitutes the optimal size is uncertain. Dalkey (1969) suggested a minimum panel of 15 to 20 to achieve reasonable accuracy from Delphi forecasts, and Yeong et al. (1989) agreed that a panel of that size is generally sufficient. Müller (1998) suggested a panel size of 40 as a general rule.

In addition, some scholars have stated that panel size is not considered a critical issue. For example, Smith (1995) argued that panel size should be determined by the number of experts available, which is typically around 40 to 50. Shafer, Moeller, and Getty (1974) illustrated that the absolute number of participants in a panel does not determine the quality of a study’s findings but the balance of expertise represented on that panel does. It has been concluded that a “balanced” panel in terms of the background and capabilities of its members should be used throughout the successive rounds of a complete Delphi study (Wheeller, Hart, & Whysall, 1990). If a panel is unbalanced, its group judgment tends to be biased in favour of the individuals who have the characteristics that are overrepresented in the panel (Garrod & Fyall, 2005).

The challenge in constructing a heterogeneous Delphi panel is to ensure that “the panel includes a diversity of cultural backgrounds, perspectives, and experience”

(Donohoe, 2011a, p. 8). McCleary and Whitney (1994) considered that a balanced panel should consist of at least 10 panellists from each industry and academic group. Seely, Iglarsh, and Edgell (1980) found that panel size is significantly influenced by the types of results desired, the scope of the exercise under question, the resources available to carry out the research project, and the time available for the completion of the project. Sadi and Henderson (2005) indicated that the optimal panel size depends on the nature, scope, and importance of a study as well as the level of knowledge and expertise of the participants. Linstone (1978) found that accuracy deteriorates with smaller panel size and improves with larger panel size. Indeed, larger groups can provide more intellectual resources than smaller ones, potentially bringing more knowledge and a broader range of perspectives to bear on a problem, but they also make conflicts, irrelevant arguments, and information overload more likely. Armstrong (1985) and Rowe and Wright (2001) both suggested that groups in general should probably comprise between 5 to 20 experts with disparate domain knowledge. Taylor and Judd (1994) asserted that a panel size range of 10 to 15 would be appropriate for homogeneous panel members (e.g. mostly technical members), whereas the size would need to be increased to 20 or 30 if the panel members are basically heterogeneous (e.g. a broad representation).

Another challenge in building an effective panel size is related to attrition rates (Donohoe & Needham, 2009). It is a common problem to have experts quit a panel during the convergence stage over rounds. Among the 30 studies examined, 26.7 per cent had an attrition rate lower than 10 per cent between the first and second rounds, 16.7 per cent had an attrition rate within the range of 10 to 20 per cent, and about 43.3 per cent had a more than 20 per cent but less than 40 per cent of their participants drop out (see Table 2.4).

Table 2.4 Attrition rates of 30 Delphi studies from Table 2.3

Group	<10%	10-20%	20-30%	30-40%	40-50%	50-60%
Attrition rate (%)	26.7	16.7	20.0	23.3	10.0	3.3

The effect of experts dropping out of successive rounds may make the conclusions misleading. For example, Murra (1979) showed that panellists who choose not to participate further are probably those who disagree most strongly with the growing conclusions of the panel. Adding new members who have not participated in previous rounds to replace members who have withdrawn is not recommended because changing membership presents a serious problem for Delphi administrators, who require stability to be maintained in order to achieve the desired outcome; it may also lead to unknown results (Murray, 1979; Donohoe & Needham, 2009). Murray (1979) alleged that Delphi results are suspicious if different panels (i.e. panels with one or more replaced members) are used over different rounds because this damages the “very core of the Delphi procedure” (p. 155). Thus, the results of Austin, Leeb, and Getzb’s (2008) study appear to be questionable as the number of panellists for rounds one through four were respectively 25, 24, 25, and 24 in their study. Spenceley (2008) used different panels for three rounds and finally had a panel of 42, 184, and 280 for three rounds; the final round included all of the participants from rounds 1 and 2 and nonrespondents to round 2 with whom contact had been confirmed. Spenceley believed that the validity of the study was not undermined as a result of increasing the panel size between rounds 1 and 2 because “it was only during Round 2 that consensus was sought” (p. 191).

To achieve stability, it is suggested that Delphi administrators should (a) develop an initial expert sample list reflecting the predetermined expert selection criteria, (b) determine a minimum requisite panel size, (c) develop a panel management plan, and (d) assess panel stability periodically throughout the process using a quality control measure based on predetermined panel membership selection criteria (Donohoe & Needham, 2009). As recommended by Green, Hunter, and Moore (1990a), the number of panellists selected at the beginning should be higher than the minimum target of 20 experts due to the expected attrition. Pan et al. (1995) advocated that "the sample size should be as large as possible to allow for subsequent drop-outs, yet small enough to ensure the respondents are all experts in their fields" (p. 32).

(4) Delphi consensus and iteration

Consensus, or convergence, is referred to as "the point at which the distribution of responses begins to stabilize" (Moeller & Shafer, 1983, p. 100). As shown in Table 2.3, most of the Delphi applications in tourism have evaluated consensus through two approaches: descriptive statistics and statistical tests. It is common to use the mean, the median and interquartiles to measure the control tendency and the standard deviation to measure the degree of convergence.

To achieve stability, McCleary and Whitney (1994) used the criterion that "no respondent can fall more than a half point above or below the mean of the responses on the nine-point scale" (p. 244). The length of the interquartile bars was also used to indicate the degree of consensus among experts in Shafer, Moeller, and Getty's (1974) study, while the median projection was represented by the peak of each bar. The evolution of consensus can also be ascertained by the descriptive analysis of group trends such as an increase in agreement percentage (Austin, Leeb, & Getzb, 2008; Donohoe, 2011b; Weber & Ladkin, 2003) and a decrease in the number of comments made (Pan, Vega, Vella, Archer, & Parlett, 1995). Some studies have suggested that agreement among 60% of a panel can be viewed as group consensus (Hill & Fowles, 1975), while others have argued that the interquartile range should be no more than 10% higher or lower than the median (Frechtling, 2001). Ulschak (1983, cited in Hsu & Sandford, 2007) recommended that consensus is achieved by having 80% of subjects' votes fall within two categories on a 7-point scale. Miller (2001) terminated a Delphi survey when he found no significant movement in the mean scores from round one to round two, but he did not define what constituted a significant change. Green (1982, cited in Hsu & Sandford, 2007) suggested that for a consensus to be achieved, at least 70 per cent of Delphi subjects needed to rate three or higher on a 4-point scale and the median had to be at 3.25 or higher.

Examples of statistical testing for consensus include the Chi-square test (Spenceley, 2008), the coefficient of variation (Lloyd, La Lopa, & Braunlich, 2000), Spearman's rank correlation coefficient (Garrod & Fyall, 2005), and Wilcoxon rank sum tests (Liu, 1988). Some studies have used a combination of two approaches; for example, Katsura and Sheldon (2008) used the median to measure consensus and also implemented a sample pair *t*-test to check the stability of the consensus between rounds at the 5% significance level.

Dyck and Emery (1970) recommended the use of a graphical presentation of the development of a consensus which not only helps the Delphi moderator to manage the operation of the rounds but also enables participants to better locate their individual views within the consensus. However, all of the aforementioned methods of determining when to consider the Delphi process complete are clearly arbitrary: There is no good reason why different rules of thumb would not be equally valid.

The Delphi process ceases when sufficient convergence is achieved; then, the group judgment is applied to inform the final report and the problem(s) being addressed (Garrod & Fyall, 2005). Some have argued that time or budgetary limitations should determine the number of rounds (Garrod & Fyall, 2005). Frenchtling (2001) set two rules on how to decide the number of iterations in the Delphi method. He stated that the most severe rule is to continue the process until there is no significant change in the median or in the interquartile range from the penultimate round to the last one, while the less severe rule is to continue the process until the interquartile range becomes relatively narrow around the median, say no more than 10 per cent higher or lower than the median. However, it should be noted that some Delphi applications do not seek convergence, divergence in group opinions being regarded as equally valid and treated as such (Garrod & Fyall, 2005). Table 2.3 reveals that the number of rounds varies from one to four but is most commonly restricted to two or three. Moutinho and Witt (1995) and Pan et al. (1995), however, identified Delphi studies with only one round.

(5) Analysis of results

Both quantitative (e.g. descriptive statistics, ratings, rankings, and statistical tests) and qualitative (e.g. extraction of themes) analyses have been employed to present Delphi results (see Table 2.3). The easiest way to present results is to rank the data; examples of studies using this form of presentation include Hawkins, Shafer, and Rovelstad (1980), McCubbrey (1999), and Lloyd, La Lopa, and Braunlich (2000). Kaynak and Macauley (1984) described a two-round Delphi study conducted among 150 panellists from different sectors of the tourism industry whose activities were closely linked to tourism and hospitality in Nova Scotia, Canada. They summarized the means, medians, standard deviations, and modes as well as each factor’s impact on tourism (e.g. median) from a pool of expert opinions. Liu

(1988) introduced the Wilcoxon rank sum test to examine whether there were significant differences between the results from a two-round survey of local experts and a one-round survey of outside experts. A Mann-Whitney test was applied by Yeong et al. (1989) to identify the difference between the perceptions of two panels based on the degree of importance they assigned to different event statements to project the future scenario of Singapore's tourism industry. There is also a need to analyse qualitative feedback (i.e. comments from panellists) so that the biases of participants and researchers can be properly acknowledged. It is worth noting that the difficulty with this is that there are very few analytic tools available for processing the large number of "non-numerical, unstructured, and rich data sets that can be captured" in Delphi studies (Day & Bobeva, 2005, p. 112).

(6) Accuracy

An apparent indicator to demonstrate Delphi's value as a forecasting tool is its accuracy. Linstone and Turoff (2002) observed that long-range forecasts tended to be pessimistic while short-range forecasts tended to be optimistic. It is very difficult to evaluate the accuracy of Delphi because the technique is based on determining the opinion of panel members and therefore the findings can be seriously affected by the possible influence of person- and situation-specific biases (Woudenberg, 1991). Furthermore, each application of the Delphi procedure is different, preventing the further possibility of comparison and measurement. This is particularly true for Delphi applications in tourism, where the Delphi technique is commonly used in forecasting the occurrence of events, identifying key issues, exposing assumptions, establishing frameworks, and constructing concepts/definitions in a particular subject/area. Apparently, the concept of accuracy does not apply to these cases.

In the general Delphi forecasting literature, there are two different views about the accuracy assessment of Delphi applications. On the one hand, Woudenberg (1991) advocated that the most feasible way of evaluating the accuracy of Delphi is by comparing it “directly to other judgment methods in the same situation” (p. 134). He summarized the following points from the previous literature: (a) a statistical aggregate of several individual judgments is more accurate than the judgment of a random individual; (b) judgment resulting from interacting groups are more accurate than statistically aggregated judgments; and (c) unstructured, direct interaction of judgments could lead to suboptimal accuracy of judgments. Rowe and Wright (1999) provided another review of the accuracy assessment of the Delphi technique compared to other group judgment methods. They found that Delphi groups outperformed statistical groups and standard interacting groups, but they did not find consistent evidence that the Delphi technique was superior to other structured group procedures. On the other hand, Martino (1970) argued that asking “How accurate is a Delphi forecast?” is a false question; instead, he asserted that a Delphi forecast should be judged in terms of its usefulness to a decision-maker rather than its accuracy. As addressed by Woudenberg (1991), some studies wrongly evaluated the accuracy of the Delphi by inferring from a few criteria, such as consensus, the lognormality of first estimates, and the relation between remoteness and the precision of a forecast.

A number of empirical studies have revealed that Delphi outperforms both statistical groups (i.e. average individual estimates without interaction; a noninteracting statistical group is also regarded to be equivalent to the first round of a Delphi poll) and interacting groups, although there is inconsistent evidence that it yields higher accuracy than a variety of other structured interacting and nominal group techniques (Wright & Goodwin, 1998). Furthermore, improvements in accuracy within the Delphi procedure have occasionally been examined by comparing the final round aggregated to the group's "best member" (an important benchmark since performance above this level would indicate process gain), but the results are not unequivocal.

Only a handful of tourism studies have investigated the accuracy of Delphi forecasts. McCubbrey (1999) predicted the impact of Internet technologies on traditional travel agents in the air travel distribution industry of the USA. In 2005, McCubbrey and Taylor (2005) compared a panel's predictions with actual results and found that the expert forecasts were very close to what actually occurred by the end of 2002. Lee and Kim (1998) used Delphi to predict the short-term effects of the 2002 World Cup on inbound tourism demand in South Korea. The panel estimated 456,000 attendees at the games, a figure which was slightly higher than the actual number of tourist arrivals (403,000) (Lee, Song, & Mjelde, 2008). Another evaluative study has been provided by Tolley, Lumsdon, and Bickerstaff (2010) who revisited the forecasting results of their Delphi study, which was conducted in 2001and predicted trends for walking in Europe by 2010. After 10 years, it was found that the expert predictions were correct in many ways; for example, as predicted, people in Europe did less walking for leisure and health in the context of rapidly rising motorization (Tolley, Lumsdon, & Bickerstaff, 2010).

(7) Evaluation of Delphi results

Day and Bobeva (2005) asserted that, to evaluate the quality of Delphi findings, the trustworthiness criteria of confirmability, credibility, transferability and dependability could complement or replace the positivist criteria of objectivity, internal validity, external validity and reliability, respectively. This is because the essence of Delphi studies can include both positivist (quantitative) and interpretative (qualitative) elements. Day and Bobeva (2005) suggested that it is more suitable and useful to examine Delphi results for their coherency, relevance, and plausibility from a qualitative perspective and to identify the explicit limitations in terms of the transferability of the results to other contexts.

The issues of validity and reliability have been subject to much more discussion than other properties. The reliability of Delphi results depends to a large extent on the expertise of the panellists, who should have high expertise in their fields (Archer, 1976). However, in practice, this is not always the case as it is likely that very few panel members will only possess true expertise in more than a limited range of a subject area. As explained by Archer (1976), the reason for this may lie in the willingness of experts to abandon their previous estimates that were unsupported by personal hard-core research or first-hand practical knowledge in favour of more popular ones nearer the median. Loo (2002) argued that another reason why the reliability of Delphi measures is challenged is because responses from different panellists to the same question could substantially differ from each other, and the final consensus reached might be due more to some pressure to confirm than to a genuine converging consensus of opinions. In response to the above criticisms, Loo (2002) suggested the use of small, nonrandom samples for a Delphi procedure; this approach could be very useful if the researcher carefully determines the key criteria for selecting panels and decides the sample size based on the expected variation in response. In addition, Loo (2002) stated that “one should not necessarily expect to achieve consensus or a decision” when doing a policy Delphi because some conflicting policy directions might emerge as a result of consensus, but such results "would not necessarily mean that the study lacks reliability or is not valid" (p. 767).

Several tourism studies have suggested that pretesting is an important way to ensure reliability for the Delphi method: examples include Lee and Kim (1998), Kaynak and Macaulay (1984), Green, Hunter, and Moore (1990a), Green and Hunter (1992), McCleary and Whitney (1994), Müller (1998), Tolley, Lumsdon, and Bickerstaff (2001), Weber and Ladkin (2003), Kaynak and Cavlek (2007), Katsura and Sheldon (2008), and Spenceley (2008). However, test-retest reliability is irrelevant since researchers expect respondents to revise their responses (Okoli & Pawlowski, 2004).

Lincoln and Guba (1985) showed that the replication of outcomes from another context is an acid test for external validity but is not meaningful for Delphi studies. Gordan (1994) explained that as the number of participants on a Delphi panel was usually small, Delphi did not (and was not intended to) produce statistically significant results. In other words, the results provided by one Delphi panel would not predict the response of a larger population or even a different panel; instead, they can only represent "the synthesis of opinion of the particular group" (Gordon, 1994, p. 4). Loo (2002) recommended that researchers use a triangulation of methods to make themselves more confident of their findings and recommendations; for example, researchers may find the combination of a Delphi study and a survey with two independent samples useful and practical for many situations (Loo, 2002).

2.4.3 Integrating forecasting in tourism[4]

Although combined forecasting has attracted broad attention in the general forecasting literature, it has yet to receive serious attention in tourism forecasting.

[4] Parts of this subsection was published in Lin (2013).

Tourism researchers have focused mostly on objectively combining two or more quantitative models using a weighting scheme, where the combination occurs within time-series methods or econometric methods or both (Shen, Li, & Song, 2008, 2011; Song et al., 2009; Wong, Song, Witt, & Wu, 2007). However, the empirical results are not yet as satisfactory as one would expect. Combined forecasts only outperform the least accurate individual forecasts; they are not as accurate as the best individual forecasts. In Wong et al.'s (2007) study, forecasting integration within the statistical category was shown to exhibit limited accuracy improvement. One common observation from the aforementioned combined studies in tourism is that none of them has incorporated contextual information into their final forecasts, which is probably the reason why the combined results are less satisfactory than expected.

A second type of combined approach is to integrate quantitative forecasts with qualitative methods; this not only forecasts tourism demand on the basis of historical data but also considers the impact of future events on tourism demand. Archer (1980) identified the need to integrate judgment and rigorous quantitative analysis. Frechtling (2001) stated that such a combination was an especially effective way of achieving convergent validity. This integrative approach has also been recommended for long-term forecasting conditions (Archer, 1980; Uysal & Crompton, 1985). Ng (1984) proposed a model consisting of three components based on different forecasting horizons — multiple regression models for long-term forecasting, the subjective-objective qualitative forecasting method for short-term forecasting, and the Delphi method for medium- and long-term forecasting — to predict and estimate the demand for leisure service manpower. However, though Ng further demonstrated that these three components supplemented and complemented one another, he did not describe how to integrate any of them. Faulkner and Valerio

(1995) illustrated an integrative approach developed by the Australian Tourist Commission and recommended that a combination of different forecasting techniques should be applied to facilitate a more meaningful dialogue between tourism analysts and decision-makers.

In the tourism integrative studies, forecasts from extrapolation methods or regression analysis have often been combined with Delphi estimates. Tideswell, Mules, and Faulkner (2001) adopted an integrative forecasting model combining quantitative methods (e.g. a Naïve model and a single exponential smoothing method) and qualitative techniques (e.g. Delphi) to measure the domestic and international tourism potential of South Australia and their study succeeded in validating forecasts for both types of tourists.

In terms of estimating tourism demand, a combination of various quantitative methods and a Delphi method is believed to generate the most reliable demand forecasts in any given situation (Goeldner & Ritchie, 2005). For example, Edgell, Seely, and Iglarsh (1980) conducted a two-stage study to combine time-series forecasts with a Delphi-type interview to forecast international tourism to the USA. Similarly, Lee and Kim (1998) employed a two-stage integrative forecasting framework to predict the international tourism demand in 2002 for the World Cup in South Korea. In the first stage, a combined time series model (time-trend regression model with an autoregressive model) was applied to forecast the number of international tourists for the tournament; in the subsequent stage, the Delphi method was used to forecast the number of the World-Cup related international tourists, non-World-Cup related international tourists, and total international tourists during the tournament. The total international tourism demand was then forecast using the combined approach (combining time-series forecasts from the first stage with the

Delphi estimates from the second stage), which predicted a figure of 5,565,757 foreign tourists for the 2002 World Cup, slightly higher than the figure on actual number of arrivals of (5,347,468) released by the Korea Tourism Organization. It is worth noting that the MAPE reported in their study was inappropriately interpreted as an indicator to conclude that the time-series model they used was highly accurate. This was because only *ex post* forecasts were made by the combined time-series model and the MAPE (5%) was calculated by comparing the difference between the fitted values and actual arrivals. Thus, the low value of MAPE in their study at most suggests a good-fit model, but it does not necessarily suggest that the model has high forecasting ability.

In addition, tourism researchers have also attempted to combine quantitative forecasts with other judgmental methods. For example, to predict leisure patterns in the UK and recreation trends in the USA, Martin and Mason (1998) adopted a combination of time series, cross-sectional, and scenario-writing techniques. Kelly and Warnick (1999) used cross-sectional cohort methods, time-series models, and consideration of trends to predict lifestyles and leisure styles, while Faulkner and Valerio (1995) offered an example of combining forecasts from econometric models with a consultative workshop.

Combined forecasting is not limited to integrating different types of forecasts obtained from different forecasting techniques. It is also used for forecasts collected from multiple sources (e.g. surveys or interviews). To predict the number of visitors to Greenwich, UK, in the pre-millennial event phase, data were collected from various sources, including visitor surveys and counts and visitor interviews at key nodal points and observations, and then combined to produce the final forecasts (Evans, 1995). International tourists and local tourist arrivals forecasts were made for paying attractions in the town of Greenwich and Greenwich Park in 1994. Lee, Song, and Mjelde (2008) used the historical data and willingness-to-visit (WTV) survey data to predict the number of domestic and international tourists to an international expo to be held in Korea in 2012. However, the lack of tests on accuracy limited the authors' findings to an appreciation that the Delphi panel predicted lower demand for the expo than the combined quantitative techniques.

Relatively little research, however, has examined the effectiveness of integrating judgmental and statistical forecasting methods in the tourism context. One notably successful application of integrative forecasting techniques has been implemented by the FAA (2010, 2011). Forecasts of aviation demand and activity measures are first made by econometric and time-series models; these are then adjusted on the basis of "expert industry opinion" to arrive at subsequent forecasts for use in making decisions. The FAA periodically reviews and adjusts its projections on the basis of forecasts and discussions with analysts outside the FAA (FAA, 2004); for example, it frequently organizes workshops to improve the reliability and utility of forecasting results. Between 1995 and 2005, the average errors and the mean absolute errors for all of the forecasts provided by the FAA were less than 2.5 per cent, suggesting significantly high forecast accuracy. Even with the negative impact of unanticipated external events (e.g. the 9/11 attacks in 2001, the outbreak of the SARS epidemic in 2003, the rapid rise of oil prices in 2004–2005), the mean absolute errors for all forecasts over the period 2002–2005, which were published a year in advance, ranged from 1.3 to 3.3 per cent, which still suggests excellent forecasting performance using the combined forecasting approach.

2.4.4 Forecasting support system with judgmental forecasting[5]

With advances in information technology, research using a forecasting decision support system (FDSS) or forecasting support system (FSS), stimulated by the rapid development of a decision support system (DSS), is becoming increasingly popular. A DDS effectively makes use of decision-making efficiency in the forecasting process achieved by combining raw data, personal knowledge, or quantitative models and identifying and solving problems in a human-machine interactive manner. Croce and Wöber (2011) emphasized that the use of an FSS is particularly meaningful in the following four situations: (a) facilitating access to data relevant for forecasts; (b) enabling selection among a set of quantitative techniques suitable for forecasting variables of interest; (c) allowing for the storage of judgmental forecasts or the adjustment of the outcome of quantitative forecasting models; and (d) providing feedback on forecasting performance (accuracy).

Implementing an FSS specific to the tourism industry would certainly provide the scope needed to gain deeper knowledge across several disciplines. This system takes advantage of statistical forecasts and the unique ability of human judgment to deal with systematic changes in patterns or relationships. One example is provided by Song, Witt, and Zhang (2008), who designed and developed a Web-based tourism demand forecasting system (TDFS). The TDFS not only utilized advanced econometric forecasting techniques for tourism demand but also incorporated the real-time judgmental contribution of experts. Furthermore, this system allowed users to perform scenario analysis or to make their own "what-if" forecasts, which incorporated uncertainty by including alternative future values of the influencing factors. More recently, Song, Lin, and Gao (2012) further developed the TDFS and

[5] Parts of this subsection was published in Lin (2013).

showed that the combination of quantitative and judgmental forecasts improve the overall forecast accuracy.

Different from the above two studies, Croce and Wöber (2011) described a group forecasting system in which base forecasts are produced by simple extrapolation forecasting methods. The system is embedded in TourMIS (a marketing-information system for the tourism industry), which supports collaborative short-term forecasting tasks among tourism managers. Estimates in TourMIS can be made either through pure judgment, one of the two established quantitative methods (i.e. Naive 2 and Winters' exponential smoothing), or a combination of the two approaches. One of the strengths of TourMIS over other tourism forecasting systems is that it can evaluate users' forecasting performance on the basis of accuracy (measured by MAPE) and reliability (defined as the capability of the user to provide accurate forecasts in the past). Croce and Wöber (2011) concluded that users' past forecasting performance can be used as a consistent indicator of expertise and utilized to qualify a system's users as reliable experts.

The studies reviewed earlier in this section provided only one direct approach to adjusting demand forecasts. Ghalia and Wang's (2000) study enriched the existing literature by proposing an intelligent system (IS-JFK) that supports two approaches to aid hotel managers in making their forecast adjustments — a direct approach and an approach via fuzzy intervention analysis. IS-JFK was designed to support the judgmental forecasting and knowledge of hotel managers. The system allows managers to adjust demand forecasts for future arrival days when there are discontinuous changes in the business environment whose impact statistical forecasting methods fail to capture. Ghalia and Wang (2000) used problem scenarios and simulation results based on actual hotel data to illustrate the effectiveness of IS-

JFK. They also addressed the importance of the cooperation of hotel managers in all aspects of conceptualizing and developing the intelligent system since their input was important in defining and characterizing the fuzzy sets used in the system.

2.5 Strategies for Improving Forecast Accuracy

The existing literature has suggested that integration leads to an increase in accuracy particularly when forecasters possess relevant domain knowledge. However, forecasting integration cannot always guarantee improvements in accuracy; sometimes, it may reduce the accuracy. A number of strategies have been proposed and investigated to facilitate the integration process, such as increasing forecasters' technical knowledge (e.g. experience, contextual information, and motivation), identifying the data characteristics (e.g. trend, seasonality, noise, or randomness; instability or discontinuities; the number of historical data points; and length of forecasting horizon), improving the format of task presentation, providing feedback (e.g. simple outcome feedback, performance feedback, task feedback, and task feedback with cognitive information feedback), using decomposition, combining forecasts mechanically, providing incentives, and using a group of forecasters (Goodwin & Wright, 1993, 1994; Lawrence et al., 2006; Remus, O'Connor, & Griggs, 1998; Webby, O'Connor, & Edmundson, 2005).

The formal interaction of judgment and statistical models requires associated techniques of modelling judgment. The simplest way is to use an arithmetic technique. MacGregor, Lichtenstein, and Slovic's (1988) study showed improved performance in terms of both accuracy and consistency across subjects with the increasing structure of the aid, but they also found that experts often make arithmetic errors when conducting judgmental decomposition. Another way to structure judgment is to organize the data presentation format for forecasters/experts, either in graphical or tabular format. There is mixed evidence about the relative merits of graphical and tabular displays, and it seems that graphics do not always help to improve forecast accuracy; they only help under certain conditions. Some tentative evidence supports the view that graphical aids lead to more accurate short-term extrapolations, while tabular aids may be superior for longer forecast lead times (Lawrence, Edmundson, & O'Connor, 1985, 1986). Remus (1984) found that when the erratic components of decisions are reduced, the tabular format outperforms the graphical format. Another study by Remus (1987) concluded that tabular aids outperform graphical aids in environments with low complexity, replicating an earlier study; however, in intermediate complexity environments, graphical aids outperform tabular aids. Benbasat and Dexter (1985) found no performance differences between subjects who used tabular reports and those who used graphical reports. It is difficult to distinguish the relative effectiveness of using graphical or tabular presentation in time-series forecasting. Lawrence et al.'s (1985) study showed no significant difference between the accuracy of judgmental forecasting made with graphical aids versus tabular aids, although they suggested that forecasts aided by a tabular presentation are more "robust".

The proposition that the decomposition of an extrapolation task improves judgmental performance has been supported by a number of pieces of circumstantial evidence. Armstrong et al. (1975) found that 12 of the 13 responses to almanac-type problems in their study were improved by decomposing the decision. This finding was not fully supported by Lyness and Cornelius (1982), who used a factorial design to compare three judgment strategies and concluded that holistic judgment could be as effective as the decomposed judgmental approach except in complex circumstances. They further revealed in their study that the algorithmic synthesis of the decomposed judgment outperformed judgmental synthesis. Given that decomposition as a strategy may lead to improved extrapolation, it is pertinent to consider how well the subtasks in a decision would be performed judgmentally. With the assistance of an interactive graphical tool called GRAFFECT, Edmundson (1990) decomposed forecasting tasks into the classical components of trend, seasonality, and randomness. This computer-aided approach supported the judgmental estimation of the trend and seasonal components and allowed the direct entry of the impact of the contextual data into the deseasonlized forecasts. This study showed a significant improvement in forecast accuracy over unaided judgments, resulting in subjective extrapolation that is superior to statistical methods alone.

More elaborate structures, such as the use of hierarchical inference, influence diagrams, scenario decomposition, system dynamics, and expert systems (Bunn & Wright, 1991), have also evolved over the past decades to explicitly structure an essentially subjective forecast. By using an expert system, analysts and forecasters attempt to replicate the procedures an expert uses to make forecasts (Collopy & Armstrong, 1992). Special rules on accumulating knowledge about methods and the problem domain are used to represent experts' reasoning in solving problems. Collopy and Armstrong (1992) found that expert systems are more accurate than unaided judgment. They also argued that there is little evidence to support the view that expert systems outperform econometric models.

Judgmental bootstrapping, a type of expert system, translates an expert's rules into a statistical model by regressing experts' forecasts on the information that they used to make their forecasts. As indicated by Armstrong (2001c), decisions and predictions from bootstrapping models are similar to those from experts and studies in the fields of psychology, education, personnel, marketing, and finance have shown that bootstrapping forecasts are more accurate than forecasts made by experts using unaided judgment. Bootstrapping is most appropriate under complex situations, where judgments are unreliable and experts' judgments have some validity. A more comprehensive discussion about bootstrapping can be found in Armstrong (2001c).

The extent to which a structured judgment can facilitate the interaction with statistical methods relies on the level of interaction between judgment and statistical methods, the key issues related to each level and the extent to which these issues provide gateways for the incorporation of judgment (Bunn & Wright, 1991). Empirical research has shown that group judgments are often suboptimal as a consequence of a number of processes related to the interactions of group members. Asch's (2003) experiment and Janis and Mann's (1979, cited in Goodwin, 2002) study showed how group pressures distort the judgment of individual group members. What is worse, people appear to be overconfident about the accuracy of their judgmental forecasts relative to that of statistical forecasts even if the evidence of the inaccuracy of judgmental forecasts is convincing (Lim & O'Connor, 1995). One way around such problems is to use structured group procedures such as the Delphi technique.

In order to integrate human judgment and quantitative analysis, the role of each step of the forecasting process should be carefully examined and investigated to take advantage of the strengths of both processes so as to minimize the potential biases. Stewart and Lusk (1994) proposed a seven-component framework for improving judgmental forecasting skills in the forecasting process. Extended from this framework, several principles developed by prior researchers are summarized in Table 2.5. A set of useful strategies can be applied to structure judgment and yield improvement in reliability, which will lead to higher forecast accuracy.

Table 2.5 Summary of strategies for improving forecast accuracy and principles for application

Method for forecasts	Component skills							Principles and conditions	Studies
	1	2	3	4	5	6	7		
Identify new descriptors through research	✓							This skill is determined by the forecast domain, the information available relevant to the forecast, and the information system, which are beyond the control of a forecaster, at least in the short run.	Stewart & Lusk (1994)
Develop better measures of true descriptors		✓							Stewart & Lusk (1994)
Train forecasters about environmental system			✓						Stewart & Lusk (1994)
Experience with forecasting problems			✓			✓	✓		Stewart & Lusk (1994)
Cognitive feedback			✓					Different types of feedback can be provided: (1) simple outcome feedback (the subjects receive the actual value after making a forecast); (2) performance outcome feedback (the subjects are provided with an error measure (e.g. MAPE) after making a forecast); (3) task feedback (consisting of information on the structure of the data of interest, e.g. upward or downward trend); (4) task feedback and cognitive information feedback (which will tell forecasters that the time series is flat and that they are overreacting to random noise).	Fischer & Harvey (1999), Goodwin & Wright (1993), O'Connor, Remus, & Lim (2005), Remus, O'Connor, & Griggs (1998)
Train forecasters to ignore nonpredictive cues			✓						Stewart & Lusk (1994)
Develop clear definitions of cues				✓				Establish explicit and agreed criteria for applying a forecasting method.	Harvey (2001), Lusk & Hammond (1991), Lusk, Stewart, Hammond, & Potts (1990)
Train to improve cue judgments				✓					Stewart & Lusk (1994)

Table 2.5 Summary of strategies for improving forecast accuracy and principles for application (Continued)

Methods for forecasts	Component skills							Principles and conditions	Studies
	1	2	3	4	5	6	7		
Improve information display				✓				(1) Organize and present information that clearly emphasizes relevant information, for example, a) use unambiguous information displays, b) avoid displays that require recognition of complex patterns or mental aggregation of many numbers to obtain a cue, and c) avoid reliance on short-term memory; (2) Use tabular form for micro-economic data, long-term forecasts, and series without trends; (3) Use graphical form for macro-economic data, short-term forecasts, and series with trends; (4) Use graphical form when making judgmental forecasts; (5) Draw a best-fitting line through the data series when making judgmental forecasts from a graphical display.	Goodwin & Wright (1993), Harvey (2001), Lawrence, Goodwin, O'Connor, & Önkal (2006), Stewart (2001)
Bootstrapping: replace a forecaster with a model					✓			(1) Include all variables that the experts may use; (2) Quantify causal variables; (3) Use the most successful experts; (4) Ensure that the variables used are valid; (5) Use a group of experts; (6) Use experts who have different backgrounds; (7) Use a large enough sample of stimulus cases; (8) Use stimulus cases that cover most reasonable possibilities; (9) Use stimulus cases that display low intercorrelations yet are realistic; (10) Use simple analysis to represent behaviour; (11) Conduct formal monitoring; (12) Use bootstrapping over judgment when a) problem is somewhat complex, b) reliable estimates can be obtained for the bootstrapping model, c) valid relationships are used in the model, and d) the alternative is to use unskilled individual judgments.	Armstrong (2001c)
Combine several forecasts					✓			(1) Combine several methods when one is uncertain about which method is most accurate or about the forecasting situation; (2) Combine forecasts when it is important to avoid large errors; (3) Use different data or methods; (4) Use at least five forecasts when possible; (5) Use formal procedures to combine forecasts; (6) Use equal weights unless one has strong evidence to support the unequal weighting of forecasts; (7) Use trimmed means; (8) Use track record to vary the weights if the evidence is strong; (9) Use domain knowledge to vary the weights on methods.	Armstrong (2001a)

Table 2.5 Summary of strategies for improving forecast accuracy and principles for application (Continued)

Methods for forecasts	Component skills							Principles and conditions	Studies
	1	2	3	4	5	6	7		
Use a structured group technique (e.g. Delphi technique)					✓			(1) Use heterogeneous experts with appropriate domain knowledge; (2) Use between 5-20 experts; (3) Provide mean or median estimate plus the reasons as feedback; (4) Three structured rounds are generally enough to achieve stable response in the Delphi pooling; (5) Equally weigh all experts' estimates and aggregate them to obtain the final forecasts; (6) Use the Delphi technique when a) expert judgment is necessary because statistical methods are inappropriate, b) a number of experts are available, and c) the alternative is simply to average the forecasts of several individuals or a traditional group.	Rowe (1998), Rowe & Wright (2001)
Require justifications of forecasts					✓	✓		This is likely to be most useful for tasks with low predictability because the reliability of information processing is a more significant problem for such tasks.	Stewart (2001)
Decompose forecasting tasks					✓			(1) Choose the form of decomposition (multiplicative or additive) according to the nature of the estimation problem; (2) Use decomposition when uncertainty is high, otherwise use holistic estimation; (3) Use multiple decomposition approaches to estimate component values when estimating quantities for which decomposition is appropriate; (4) Use decomposition only when one can estimate component values more accurately than the target quantity.	MacGregor (2001)
Mechanical combination of cues					✓			Use this method when (1) information can be processed mechanically without losing important cues and/or (2) the forecasting environment contains a high degree of uncertainty.	Stewart (2001)
Statistical training						✓	✓	This involves not only the forecasters' technical knowledge in areas including the characteristics of time series, model building, and statistical forecasting methods but also knowledge about the judgmental analysis of data (e.g. visual check for trends and levels) and the biases inherent in human judgment.	Goodwin & Wright (1994), Stewart & Lusk (1994)

Table 2.5 Summary of strategies for improving forecast accuracy and principles for application (Continued)

Methods for forecasts	Component skills							Principles and conditions	Studies
	1	2	3	4	5	6	7		
Feedback about nature of the bias in the forecasts						✓	✓	(1) Obtain feedback about the forecast accuracy and the reasons why errors occurred. Feedback should be explicit, systematic, and frequent; (2) Describe reasons why the forecasts might be wrong; (3) Review the forecasting methods periodically and identify the reasons for large forecast errors.	Arkes (2001), Goodwin & Wright (1994), O'Connor, Remus, & Lim (2005)
Search for discrepant information						✓			
Statistical correction for bias						✓		Use (1) when judgmental bias is of the highest concern; (2) forecasters and users are not the same; and (3) unbiased forecast is the goal.	Goodwin & Wright (1994), Stewart & Lusk (1994), Sanders & Ritzman (2004)

Note: 1: Environmental predictability, 2: Fidelity of the information system, 3: Match between environment and forecaster, 4: Reliability of information acquisition, 5: Reliability of information processing, 6: Conditional (regression) bias, 7: Unconditional (base rate) bias.
Source: Adapted from Stewart and Lusk (1994, p. 587).

2.6 Chapter Summary

There is little doubt about the critical role that judgment plays in a successful tourism forecasting process, either through a quantitative or judgmental forecasting approach. Judgment can be integrated into every stage of quantitative forecasting including selecting variables, deciding the functional form, building models, estimating parameters, and conducting data analysis. When applying a judgmental forecasting technique, judgment plays an even more important role, from selecting judges to deciding how to analyse and report final judgments.

To this point, this chapter has provided a review of studies on why, when and how to incorporate judgments into a quantitative forecasting process, and the strategies to use to produce more accurate forecasts. Overall, these studies provide strong evidence to support such integration, which implies that the appropriate integration of judgmental and statistical forecasting approaches is likely to improve forecast accuracy when performed under certain conditions; some of these conditions are identified in this study but many others remain unexplored. Some general observations are made below.

First of all, prior research on the integration of judgmental and statistical forecasts has reported mixed results in terms of either improvement, deterioration, or no difference in forecast accuracy compared to its constituent forecasts. However, most of these studies were conducted in a laboratory environment where the subjects (usually students) had minimal forecasting and/or subject-matter expertise, and the data series were typically artificially produced or taken randomly from M-competition series (Eroglu & Knemeyer, 2010). The main advantages of this method are its internal validity and the relative ease of conducting this type of study compared to fieldwork. However, the representativeness of these experiments could be doubtful and the results usually lack generalizability. In fact, some results could even be misleading. For example, earlier studies actually cautioned against the application of judgment in the forecasting process. After the 1990s, more studies were undertaken in real business situations, and these have shown that the integration approach is indeed effective in improving forecast accuracy, especially when the forecaster possesses sufficient domain knowledge that the statistical forecasting methodology ignored. Thus, there is a great need to conduct studies in realistic conditions.

Second, it is worth noting that all of these studies were conducted using different data sets, different subjects, different environmental or experimental situations, different forecasting methods, different strategies to assist the forecasting process, and different level of expertise. Integration has both positive and negative impacts on individual forecasts; thus, the final recommendations of the forecasters in these studies were based upon the average improvement or deterioration of forecast accuracy. Hence, it is important to identify the specific conditions under which the integration of two forecasts will result in higher forecast accuracy.

Third, most studies on integration are not based on theories but on empirical evidence in the previous literatures. In other words, there are no systematic theories to guide researchers and forecasters to conduct integration studies; they have to borrow relevant theories and literature from psychology, behavioural decision theory, organization behaviour, statistics, econometrics, and economics (Eroglu, 2006). Therefore, inconsistent or inconclusive and even conflicting results are reported which not only fail to generalize findings and conclusions but also fail to achieve a better understanding of the integration process. Consequently, the lack of a theoretical foundation is likely to mislead researchers and make them unable to formulate sound hypotheses, design suitable modelling and forecasting methods, and generalize their findings. Therefore, theoretically based research is needed to enhance the literature on the integrative forecasting process.

Fourth, most of the past studies on integrating judgmental and statistical forecasts have been predominantly focused on statistical extrapolation, such as Naive models, exponential smoothing methods, and Box-Jenkins time-series models, while econometric models, especially the most recently developed techniques such as ECM, TVP, and VAR, have largely been ignored. In particular, there are few studies in the tourism demand forecasting literature where research efforts have been made in terms of the combination of statistical forecasting methods. It is even rare for standard econometric techniques to be used to integrate quantitative and qualitative forecasts. Thus, the current study is believed to be the first of its kind in the tourism literature. More importantly, there is a great need for such an integration of forecasting methods given the volatile nature of the industry caused by possible environmental impacts that cannot easily be picked up by statistical modelling but can be picked up by human contextual knowledge.

Last but not least, the integration of statistical and judgmental research methods lends depth and clarity to tourism demand forecasting research. However, there are potential problems when making such attempts, and thus researchers may give up because they lack the expertise required to implement both types of methods. In addition, it is time-consuming, labour-intensive, and expensive to use multiple approaches.

A number of judgmental forecasting methods are available in tourism, but whichever technique is used, it is essential to recognize both its merits and limitations since this will affect the quality of the forecasts obtained. Choosing an appropriate forecasting method depends on multiple considerations, including the level of uncertainty involved, the level of forecast accuracy required, the availability of resources, and the time needed to obtain the forecasts. However, unlike quantitative forecasting models, it is difficult to evaluate the performance of judgmental forecasts. Several issues remain regarding the final evaluation of judgmental forecasting, such as the utility, accuracy, and reliability of judgmental forecasts and the need for validation. Many studies have also been carried out primarily for practical purposes, which adds to the difficulty in ascertaining the true utility of these forecasts. In addition, researchers have not paid much attention to revisiting their forecasts, thus missing the chance to evaluate the utility of judgmental forecasts. Reports on forecasting studies have thus required that more thorough comparisons be made among the various judgmental forecasting methods.

Depending on the target audiences involved in forecasting tasks, judgmental forecasting techniques are divided into four categories, namely asking the stakeholders, asking the experts, asking the public, and judgment-aided methods. The findings suggest that the Delphi method and scenario writing are the two most popular judgmental forecasting techniques used in tourism studies. Delphi has been widely applied in projecting potential market trends or conditions, predicting the likelihood or the time of the occurrence of specified events and their impact on tourism and forecasting tourism demand variables. Most applications of the Delphi technique, however, have been in the area of long-range forecasting. Although few studies have applied it to forecast tourism demand variables, tourism researchers have used it to produce quantitative forecasts and have integrated it as a major component in combined forecasting. Nevertheless, more attention should be placed on evaluating the performance of Delphi forecasts, especially where quantitative estimates are generated from a panel. Performance should also involve both accuracy and reliability. Subsequent studies are also needed to show the comparative accuracy of Delphi studies over other judgment methods.

It is quite difficult to capture such a diverse, dynamic, and changeable phenomenon as tourism in a limited number of variables. Sociological and psychological factors are difficult to express quantitatively, and unexpected crises and disasters are impossible to forecast. 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. It would therefore be natural 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.

To date, the combination of multiple methods is still not widely accepted as a viable research strategy in the tourism demand forecasting field. Tourism demand forecasters and practitioners have indicated that such research is necessary to develop and strengthen our understanding of many tourism-related issues, the research norms and scientific dogma regarding appropriate methods may shift to a new, more integrative paradigm. Therefore, there is a clear need to develop a research framework for the integration of statistical and judgmental forecasts in tourism demand forecasting. Full details of this study’s research design, data-collection methods, and data analysis are presented in the next chapter.

Chapter 3 : Methodology

3.1 Introduction

After reviewing studies on integrating forecasting in the general forecasting literature and the tourism literature, this chapter describes the methodological issues and decisions related to this study, aiming to establish a systematic framework for integrating judgmental and statistical forecasting methods. This chapter is divided into seven parts. In broad outline, the chapter describes the variables that were included in the statistical (econometric) models in this study, the data-collection methods, the econometric models used in modelling and forecasting, how forecasting adjustments were made via the HKTDFS, the reason for selecting Delphi as a group forecasting procedure in this study, how forecasting performance was evaluated, and the reasons for using the in-depth interview method in this study.

The chapter is structured as follows: Section 3.2 describes the research design strategy adopted in this study; Section 3.3 provides details on the variables and data sources; Sections 3.4 presents the econometric models used to make statistical forecasts in the present study; Section 3.5 starts with a brief introduction to the key features in the HKTDFS, followed by the justifications and the procedure for applying the Delphi forecasting method; Section 3.6 describes the research hypotheses in this study; Section 3.7 presents the ways to evaluate forecasting performance and the statistical tests used to examine the statistical significance of the accuracy difference; and Section 3.8 provides the justifications for selecting the in-depth interview method and presents the procedure used to conduct interviews in the present study. A short summary is provided at the end of the chapter.

3.2 Research Design

The *sequential explanatory design* strategy, which is a two-phase mixed methods design (see Figure 3.1), was adopted in this study to achieve the proposed research objectives. This design strategy is characterized by the "collection and analysis of quantitative data in a first phase of research followed by the collection and analysis of qualitative data in a second phase that builds on the results of the initial quantitative results" (Creswell, 2009, p. 211). It is typically used to explain and interpret the findings of a primarily quantitative study by collecting and analysing follow-up qualitative data (Creswell, 2009). A greater emphasis is typically placed on the quantitative (QUAN) methods rather than the qualitative (qual) methods because this research design usually begins quantitatively.

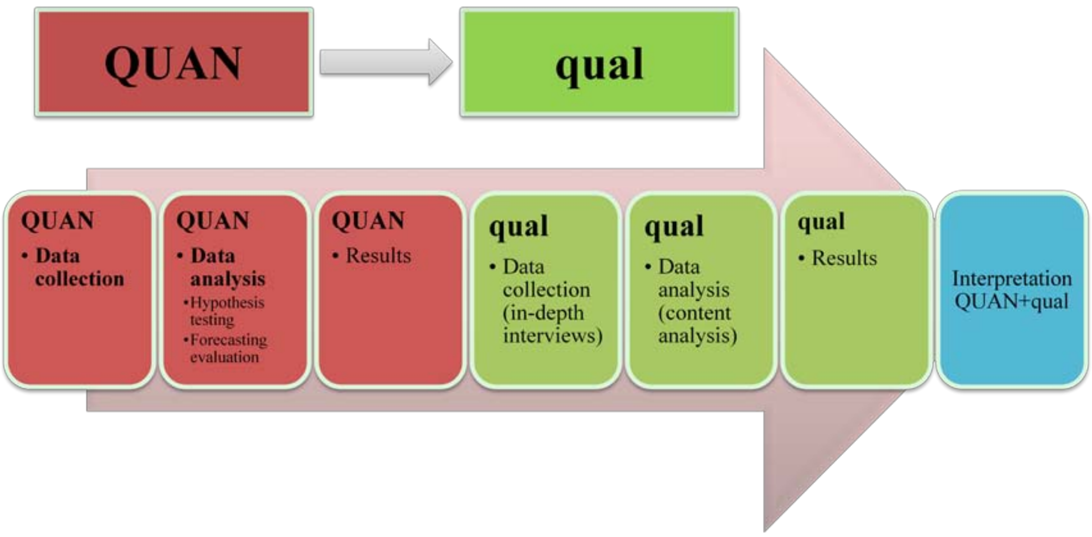


Figure 3.1 Sequential explanatory design (QUAN emphasized)

Note: QUAN and qual denote quantitative and qualitative research methods, respectively.

More specifically, the *follow-up explanations model* described by Creswell and Clark (2007) was applied in this study to explain or expand on quantitative results by using qualitative data. The use of this model allows researchers to identify "specific quantitative findings that need additional explanation, such as statistical differences among groups, individuals who scored at extreme levels, or unexpected results" and then collect "qualitative data from participants who can best help explain these findings" (Creswell & Clark, 2007, p. 72).

The advantages of the sequential explanatory design strategy include the following: (a) it does not necessarily require a specific theoretical perspective; (b) its two-phase design structure makes it easy to implement because the steps fall into clear and separate stages so that researchers can conduct the two methods separately and collect only one type of data at a time; (c) it is easy to describe and report the findings in two phases, which makes it straightforward to write and thus provides a clear delineation for readers; and (d) it appeals to quantitative researchers as it often requires a strong quantitative orientation at the beginning of the study (Creswell, 2009; Creswell & Clark, 2007). One main drawback of this design is that it requires a considerable amount of time to implement.

3.3 Variables and Data Sources

3.3.1 Tourism demand measures

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 et al., 2010; Song, Witt, & Li, 2009). In this study, the visitor arrivals variable was selected to measure inbound tourism demand in Hong Kong. In line with the definition used by HKTB, in this study, the term *visitor arrivals* refers to arrivals by all non-Hong Kong residents through immigration formalities (Census and Statistics Department, 2011). The arrivals figures by *country of residence* are based on a systematic sampling of arrivals cards, collected from the Immigration Department of Hong Kong (HKTB, 2011). The arrivals figures include both *overnight visitors* (defined as those who stay at least one night in collective or private accommodation in Hong Kong) *and same-day in-town visitors* (defined as those who pass through Hong Kong Immigration, but do not spend a night in collective or private accommodation in Hong Kong).

3.3.2 Determinants of tourism demand

According to the existing literature, the most commonly considered influencing factors of tourism demand are tourists' income, the own price of the tourism products, the price of substitute tourism products, tourism marketing expenditure, and travel costs from the origin countries/regions to the destination (Song & Li, 2008; Song et al., 2010; Song, Witt, & Li, 2009; Witt & Witt, 1995). Song and Li (2008) concluded that the income level of the origin country/region, the relative tourism prices of the destination relative to those of the origin country/region, tourism prices in competing destinations (i.e. substitute prices) and exchange rates are the most significant determinants of tourism demand. In addition, travel costs, marketing expenditure, and special events also influence tourism demand. Song and Li's (2008) review is consistent with those carried out by Witt and Witt (1995), Lim (1997, 1999, 2006), and Li, Song, and Witt (2005).

Of the aforementioned explanatory variables, tourism income is regarded as the most frequently used and most statistically significant variable. Tourism demand is also sensitive to one-off events, which can be divided into two categories according to the direction of the impact on demand: positive events (e.g. the Olympic Games, exhibitions, and visa-free arrangements) and negative events, such as man-made crises (e.g. terrorist attacks, wars, economic crises, and international trade conflicts) and natural disasters (e.g. earthquakes, hurricanes, tsunamis, floods, and epidemic diseases) (Song et al., 2010). Over the past 10 years, a number of studies have focused on the quantification of such external shocks on tourism demand. Song and Li (2008) suggested that the general procedure for such post-event analysis is to estimate a reliable model using historical data prior to the event, and then to use that model to predict tourism demand for the period affected. The difference between the actual demand as a result of the event in question and the estimated demand is then taken as the event's impact on tourism demand.

In addition, 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. The inclusion of the lagged dependent variable can be justified on two sides: demand and supply. The inclusion is based on the grounds of habit persistence. Once people have been on holiday to a particular destination and liked it, it is highly possible that they will revisit that destination. There is much less uncertainty associated with a repeat visit to that destination compared with travelling to a previously unvisited foreign country/region. Additionally, "word of mouth" recommendation may also play an important role in tourists' destination selection and maybe even more important than commercial advertising (Witt & Witt, 1992). Witt and Witt (1992) suggested that "as people are, in general, risk (i.e. uncertainty) averters, the number of people choosing a given alternative in any year depends on the numbers who chose it in previous years" (p. 24).

On the other hand, another justification for including a lagged dependent variable is to accommodate supply constraint. Supply constraints may take the form of shortages of hotel accommodation, passenger transportation capacity and trained staff, and these cannot often be increased rapidly. It requires time to build up contacts among tour operators, hotels, airlines, and travel agencies; and it is unlikely for a highly developed tourism destination to dwindle rapidly. To postulate a partial adjustment process to allow for rigidities in supply, the following equation is specified:

$$Q_t = (1-\mu)Q_{t-1} + \mu Q_t^* \tag{3.1}$$

where Q_t is the actual level of demand at time t, Q^* is the desired level of demand at time t, μ is the speed of adjustment ($0 < \mu < 1$). Here, μ lies strictly between zero and unity, indicating that there is some adjustment, but it is incomplete.

It is however important to note that there are other factors such as marketing expenditure of the tourism product/service providers (both at the destination and firm level), the change of tastes and preferences towards Hong Kong as a tourist destination in the source markets. The difficulty in accessing the relevant marketing data hinders its application in most empirical studies (Kulendran & Dwyer, 2009; Zhang, Kulendran, & Song, 2010). Moreover, previous studies such as Chon, Li, Lin, and Gao (2010), Song, Kim, and Yang (2010), Song and Lin (2010), and Song, Wong, and Chon (2003), have proved that these factors do not affect the overall goodness of fit of the models. Given the above reasons and a lack of sufficient historical data, it was thus decided that tourism marketing expenditure and travel cost variable were left out in this study. The details of the dependent and independent variables are provided in Table 3.1.

Table 3.1 Variable selection

	Formula/Description	**Sources**
Dependent Variables	Visitor arrivals VA_{it}): Visitors from the i^{th} origin country/region at time t.	HKTB
Independent Variables		
Real GDP index (Y_{it}), 2005=100		IMF
Own price (P_{it}), measured by the exchange-rate-adjusted consumer price index (*CPI*), 2005=100	$P_{it} = (CPI_t^{HK} / \ EX_t^{HK}) / (CPI_t^i / \ EX_t^i)$ at time t, where CPI_t^{HK} and are the *CPI*s for Hong Kong and the i^{th} origin country/region at time t, respectively, and and EX_t^i are the exchange rate indexes for Hong Kong and i^{th} origin country/region at time t, respectively.	IMF
Substitute price (P_{ist}), 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*, 2005=100	$P_{ist} = \sum_{j=1}^{6} (CPI_{jt} / EX_{jt}) w_{jt}^i$ (j = 1, 2, …, 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).	(1) IMF (2) Official websites of statistical bureaus or departments
Dummy variables: Seasonal dummies and dummies for one-off events (e.g. Hong Kong's return to China in 1997, Asian Financial crisis in 1997/1998, SARS in 2003, global financial/economic crisis since 2008, and outbreak of H1N1 flu, etc.), and other market-specific dummies.		

3.3.3 Data sources

The demand model 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 income variable, *Y*, measured by the real GDP index (2005=100), was 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. CPIs (2005=100) and exchange rates were also obtained from IMF. Six competitive destinations of Hong Kong, including 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. 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).

The inclusion of six markets was due to the following three considerations. First, the selected six origins occupied more than 80% of the inbound market share in Hong Kong for the year 2011: China (67.03%), Taiwan (5.13%), Japan (3.06%), the USA (2.89%), Australia (1.54%), and the UK (1.21%). Among the six selected markets, China, Taiwan, Japan, and the USA were the top four source markets in

Hong Kong. Second, such a selection provides a mixed profile of long-haul and short-haul markets with different characteristics. Third, according to the experience from the past Delphi surveys conducted by the HKTDFS, the selected six markets received the most comments from the panel due to the familiarity among the respondents. Though they were also interested in the emerging markets such as the Russian Federation and the UAE, the arrivals data for such source markets were insufficient to estimate the econometric models for those countries.

3.4 Econometric Forecasting Models[6]

A major advantage of econometric models over time-series models is that the former "explicitly take into account the impact on the variable to be forecast of changes in the determining forces, which permits a company to link its forecasting with tactical and strategic plans for the future" (Witt & Witt, 1992, p. 122). By using econometric forecasting methods, one can explore the consequences of alternative future policies on tourism demand, something that is not possible with time-series methods. The modelling procedure shown in Figure 3.2 was used in this study to estimate models, conduct diagnostic tests, and select the most appropriate functional form of the models.

[6] Parts of this subsection was published in Chon et al. (2010).

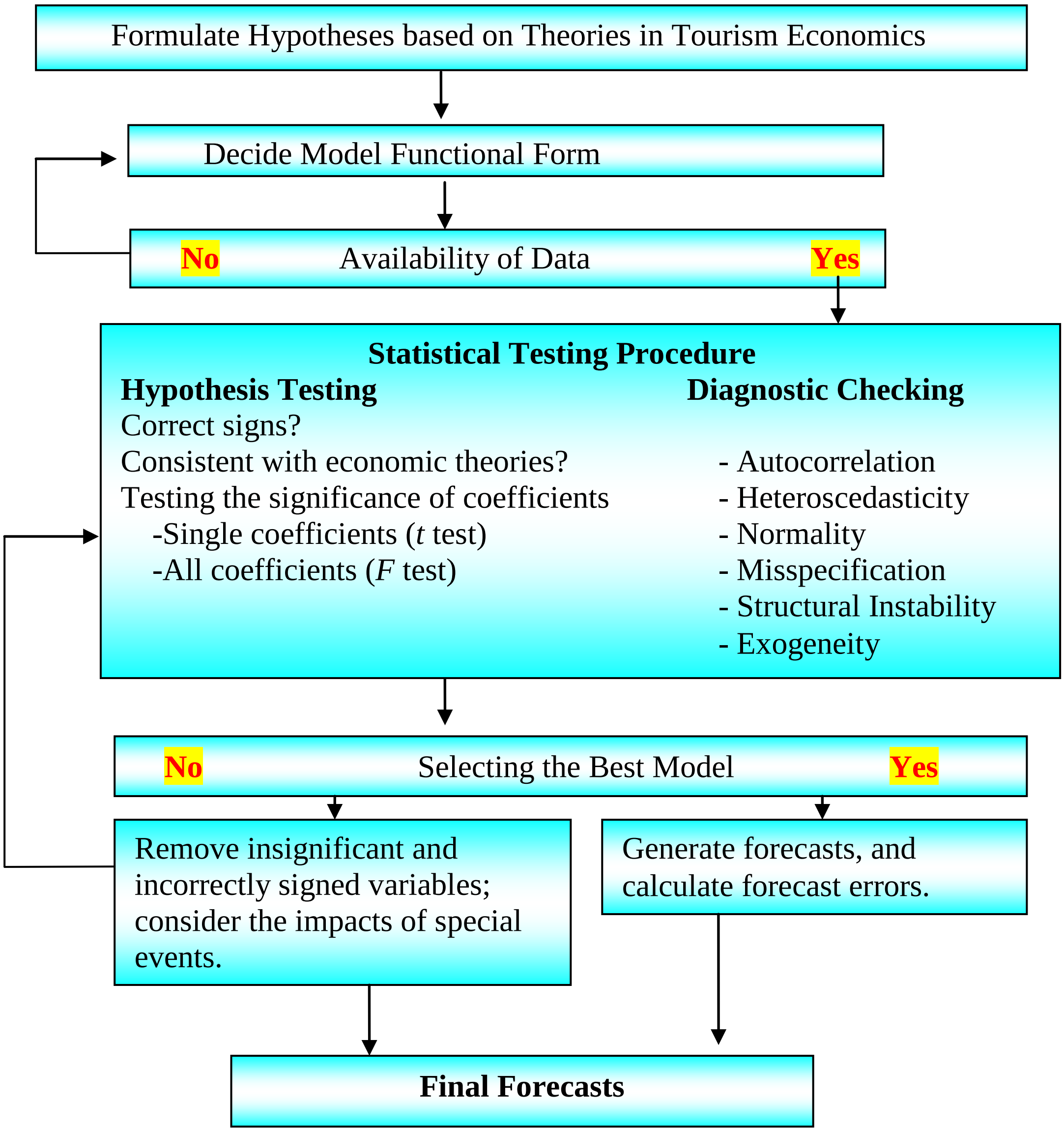


Figure 3.2 The diagram of econometric modelling and forecasting

3.4.1 Functional form

As aforementioned, the vital factors that determine tourism demand for most tourism products/services are the tourist income, tourism prices in the destinations, and one-off events. This study adopted the general mathematical notation in order to model Hong Kong's tourism demand by visitors from major source markets, which can be written as:

$$VA_{it} = AY_{it}^{\beta_1} P_{it}^{\beta_2} P_{st}^{\beta_3} e_{it} \tag{3.2}$$

where VA_{it} is the tourism demand variable measured by visitor arrivals from the i^{th} source market to Hong Kong at time t, P_{it} is the price of tourism in Hong Kong at time t relative to that in the i^{th} source market, Y_{it} is the income of tourists from the i^{th} source market at time t, P_{st} is the price of tourism in the competing destinations at time t, and e_{it} is the residual term used to account for some other economic and non-economic factors that may have been omitted for the good of the model tractability or most commonly, due to data unavailability.

The power function in Equation (3.2) is used in model estimation for the following two reasons. First, most previous empirical studies have suggested that tourism demand can better be modelled by the power function than the simple linear demand function in terms of models' statistical significance and forecasting ability (Song, Wong & Chon, 2003; Witt & Witt, 1992). Second, the power function can be transformed into a log linear specification, which can easily be estimated by Ordinary Least Squares (OLS). The estimated coefficients of the explanatory variables in the log linear model can be interpreted directly as demand elasticities.

When logarithmic transformation is carried out, Equation (3.2) is transformed into the following form:

$$\ln VA_{it} = \beta_0 + \beta_1 \ln Y_{it} + \beta_2 \ln P_{it} + \beta_3 \ln P_{st} + \varepsilon_{it} \tag{3.3}$$

where $\beta_0 = \ln A$ and $\varepsilon_{it} = \ln e_{it}$ ($\varepsilon_{it} \sim N(0,\sigma^2)$). β_1, β_2, and β_3 are income, own price, and cross price elasticities, respectively. In Equation (3.3), a positive sign is expected for income elasticity ($\beta_1 > 0$) and a negative sign for own price elasticity ($\beta_2 < 0$). The sign of β_3 is indeterminate as it depends on whether the origin market takes

Hong Kong as a competitive or complementary destination of its competitors. The sign of β_3 is therefore for empirical evidence to resolve.

However, because Equation (3.3) is a static model, it relates to the current tourism demand variable and to the current values of the influencing factors and therefore does not consider the dynamic feature of tourists' decision-making process. Song, Wong, and Chon (2003) argued that tourism demand is a dynamic process because tourists make decisions about destination choice with time leads. This means that models used for analysing and forecasting tourism demand should mirror this feature.

To comply with this requirement, the ARDL-ECM model was applied to capture the changing aspects of economic activities. In line with the majority of the tourism demand literature such as Chon et al. (2010), Song, Kim, and Yang (2010), Song and Lin (2010), and Song, Lin, Witt, and Zhang (2011), the following model was employed to model and forecast the inbound tourism demand in Hong Kong.

$$\begin{aligned}\Delta \ln VA_{it} &= \alpha_0 + \sum\nolimits_{j=1}^{p_1} \psi_{Qj} \Delta \ln VA_{i,t-j} + \sum\nolimits_{j=0}^{p_2} \psi_{Yj} \Delta \ln Y_{i,t-j} \\ &+ \sum\nolimits_{j=0}^{p_3} \psi_{Pj} \Delta \ln P_{i,t-j} + \sum\nolimits_{j=0}^{p_4} \psi_{P_{is},j} \Delta \ln P_{is,t-j} \\ &+ \pi_1 \ln VA_{i,t-1} + \pi_2 \ln Y_{i,t-1} + \pi_3 \ln P_{i,t-1} + \pi_4 \ln P_{is,t-1} \\ &+ \delta_1 D_1 + \delta_2 D_2 + \delta_3 D_3 + \sum\nolimits_{d=1}^{D} \theta_d Dummies + u_{it}\end{aligned} \tag{3.4}$$

where Δ is the first difference operator (i.e. $\Delta X_t = X_t - X_{t-1}$), 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 above equation describes the short-run dynamic interactions between the visitor arrival variable and its determinants. Equation (3.4) indicates that the demand for tourism in the current period is affected by the values of the lagged demand variable as well as the current and lagged values of the influencing factors.

This specification takes the time path of tourists' decision-making process into consideration.

Table 3.1 provides the detailed description of the definition of the income variable, own price, and substitute price variable. The six substitute/competitive destinations were chosen by considering the geographical proximity and cultural dimensions. It should be noted that once one of the six substitute markets is considered as a source market in a demand model, it is removed from the calculation of the substitute price for this model.

While estimating Equation (3.4), three seasonal dummy variables (D_1, D_2, and D_3) were included to capture the seasonality effects on visitor arrivals and one-off event dummy variables (*Dummies*) were used to capture the influences on the demand for Hong Kong inbound tourism. According to Greene (2008), a dummy variable is a variable that "takes the value of one for some observations to indicate the presence of an effect or membership in a group and zero for the remaining observations" (p. 106). The dummy variables assume a value of 1 in the respective years and quarters where they have an effect, and 0 otherwise (Hardy, 1993). In their review article, Song, Witt, and Li (2008) concluded that researchers often include dummy variables in international tourism demand models to capture the impacts of "one-off" events. A number of events were taken into consideration in this study, such as, the handover of Hong Kong to China in 1997, the SARS epidemic in 2003, the bird flu in 2003, Beijing Olympic Games in 2008, the global financial and economic crisis in 2008, and relevant country- or region-specific dummies (e.g. the 9/11 terrorist attack in the USA).

In tourism demand analysis most empirical studies have suggested that it is sufficient to set up the initial lag length of p = 4 for quarterly data and p = 1 for annual data (e.g. Song, Chon, & Wong, 2003; Song, Witt, & Li, 2009). The lag order q_i (i = 1, 2, 3, 4) in Equation (3.4) was determined by the Akaike information criterion (AIC). This study adopts the AIC as "the AIC model appears to be statistically more acceptable than the SBC criterion" (Halicioglu, 2008, p. 8). AIC is the preferred criterion to the Schwarz Bayesian Information Criterion (SBC) and the Consistent Akaike Information Criterion (CAIC) for the following two reasons. First, the complexity of the model will be penalized more heavily by SBC and CAIC than by AIC, which may lead to contradictory model selections (Song, Witt, & Li, 2009). Second, AIC tends to asymptotically perform better than SBC and CAIC in the empirical studies in terms of model collection (Anderson, Burnham, & White, 1998; Yang, 2005).

3.4.2 Testing for nonstationarity and stationarity: Unit root tests

Many economic time series (e.g. GDP, exchange rates) exhibit trending behaviour or nonstationarity in the mean. Two common trend removal or de-trending procedures are first differencing and time-trend regression (Zivot & Wang, 2006). To render the data stationary, unit root tests can be applied to determine if the trending data should first be differenced or regressed on deterministic functions of time (Zivot & Wang, 2006). If the series are found to be $I(1)$ after taking first difference, cointegration techniques can be used to model the long-run relations. However, if the series are found to be a combination of $I(0)$ and $I(1)$, conventional cointegration techniques would be inappropriate. This study thus adopted the ARDL bounds test proposed by Pesaran, Shin, and Smith (2001). One of the assumptions of the approach adopted by Pesaran, Shin, and Smith (2001) is that all variables are $I(0)$ or $I(1)$. In this study, the aim of conducting unit root tests was to ensure that none of the variables included in Equation (3.4) was integrated of order 2 or above, thus avoiding spurious regression relationships. Moreover, the computed *F*-statistics provided by Pesaran, Shin, and Smith (2001) become invalid in the presence of *I*(2) variables (Fosu & Magnus, 2006).

The Augmented Dickey-Fuller (ADF) test is used for the null hypothesis that a time series y_t is *I*(1). Stationarity tests, on the other hand, are used for the null that y_t is *I*(0). The most commonly used stationarity test is the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test developed by Kwiatkowski, Phillips, Schmidt, and Shin (1992). However, the ADF test is not generally reliable in small samples due to its poor size and power properties; for example, it tends to overreject the null hypothesis when it is true and underreject it when it is false (Harris & Sollis, 2003). Similar problems have been found with the KPSS test (Caner & Kilian, 2001).

Generally, the ADF test cannot distinguish highly persistent stationary processes from nonstationary processes very well. In addition, the ADF test that includes a constant and a trend in the test regression has less power than a test with only a constant in the test regression. To maximize power against very persistent alternatives, the Ng-Perron test proposed by Ng and Perron (2001) was used in the current study. Ng and Perron (2001) also addressed the problem of the sensitivity of unit root testing to lag choice; they proposed the modified information criteria (MIC) as a new set of information criteria.

The ADF, KPSS, and NP tests were applied in the current study to test for the presence of unit roots in the series included in Equation (3.4). Detailed technical details regarding these three tests can be found in the literature (Harris & Sollis, 2003; IHS EViews, 2009b; Zivot & Wang, 2006).

3.4.3 Testing for long-run relationships

To test for the existence of the long-run relationships between the visitor arrivals variable and its determinants, the bounds testing procedure proposed by Pesaran, Shin, and Smith (2001) was employed in this study. To implement the bounds tests, it is essential to adopt a conditional autoregressive distributed lag model (ARDL) as Equation (3.4). The π coefficients in Equation (3.4) specify the long-run relationship between the demand and its determinants. If the values of π are zero, then no long-run relationship exists. *F*-test and *t*-test are used to test for the null hypothesis of no long-run relationship against the alternative hypothesis that at least one π is non-zero.

The bounds test for examining evidence for a long-run relationship starts with the F-test. The null hypothesis of the *F*-test is H_0: $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0$ of no cointegration among the variables in Equation (3.4), against the alternative that at least one π is non-zero. Pesaran, Shin, and Smith (2001) provide the critical values of the lower and upper bounds in Tables CI (pp. 300-301) and CII (pp. 303-304). The critical values for the lower bound were obtained based on the assumption that all variables are purely $I(0)$, whilst those for the upper bound assume that all variables are purely $I(1)$. If the computed *F*-statistic lies outside the critical values of the boundaries, then a conclusive result is reached without identifying the cointegration rank. More specifically, if the computed *F*-statistic is higher than the upper bound, then the null hypothesis of no cointegration is rejected, which implies that there may be long-run relationships between the variables. If it is below the lower bound, then the null hypothesis cannot be rejected. However, if the calculated *F*-statistic lies between the upper and lower bounds, then the test results are inconclusive.

Pesaran, Shin, and Smith (2001) suggest that if the null hypothesis of the bounds test is rejected, the *t*-test should be performed to identify the cointegration relationships. The *t*-test has the null hypothesis that there is no cointegration (H_0: $\pi_1 = 0$) with respect to the lagged levels of the tourism demand in Hong Kong from origin countries. If this null hypothesis is false, then the result should exhibit a large value of *t*-statistic, at least asymptotically, thus confirming the existence of cointegration relationships among the levels of variables.

A detailed procedure is provided by Narayan and Smyth (2006) to explain if a time trend is needed for the implement of the bounds *F*-test; it is however argued by them that "in the spirit of the bounds test, model with a time trend is invalid because for the model to be valid there should be only one long-run relationship" (p. 116). This study adopts this idea to apply the approach that implicitly assumes that Equation (3.4) is free from a time trend due to the differenced variables. In other words, the *F*-test indicates that there exists only one cointegrating relationship without a trend in which the dependent variable is visitor arrivals.

Once a long-run relationship has been established, the conditioned long run models can be obtained from the reduced form of Equation (3.4) assuming the first differenced variables jointly equal zero in the long run (De Vita & Abbott, 2002; Rushdi, Kim, & Silvapulle, 2012). It can be expressed as:

$$\ln VA_{it} = \lambda_0 + \lambda_1 \ln Y_{it} + \lambda_2 \ln P_{it} + \lambda_3 \ln P_{ist} + v_{it} \tag{3.5}$$

where $\lambda_0 = -\alpha_0 / \pi_1$, $\lambda_1 = -\pi_2 / \pi_1$, $\lambda_2 = -\pi_3 / \pi_1$, and $\lambda_3 = -\pi_4 / \pi_1$.

3.4.4 Model testing procedure

In practice, not all of the variables included on the right-hand side of Equation (3.4) would be statistically significant once the model is estimated. Therefore, a testing-down procedure termed the general-to-specific approach (Hendry, 1995) was adopted in this study to determine which variables should be kept in the final model based on their statistical significance and economic acceptability. The final models were chosen on the basis of three criteria: the estimated coefficient has to have the right sign; the inclusion of a variable should be based on an economic and/or technical rationale; and the *t* statistics of the estimated coefficients should approach significance at the 10% significance level at least (Vanegas & Croes, 2000).

The test procedure began with an estimation of Equation (3.4) using OLS to check the statistical significance of all the variables. The statistically insignificant variables were then eliminated from the specification one by one in accordance with the *t* statistics of the estimated coefficients, starting with the least significant ones. It should be noted that in some cases, a variable with a nonsignificant coefficient was retained because the estimates obtained were more reasonable than those obtained when the variable was omitted (Vanegas & Croes, 2000).

According to Hendry (1980), diagnostic checking is a critical part of the whole process of model selection: "Rigorously tested models, which adequately describe the available data, encompass previous findings and were derived from well based theories would greatly enhance any claim to be scientific" (p. 403). Thus, the model is tested by using a number of diagnostic statistics to test the adequacy of the specification of a regression equation after getting rid of all of the insignificant variables and the variables with incorrect signs from the specification. Song and Witt (2000) discussed the required diagnostic statistics, which include tests for heteroscedasticity, autocorrelation, normality, and forecasting ability. In other words, the final model is required to pass a battery of diagnostic tests, including the tests for nonnormality, autocorrelation, heteroscedasticity, model misspecification, structural instability, and nonexogeneity.

Testing for nonnormality

The Jarque-Bera (J-B) test is used for testing whether residuals are normally distributed. The J-B statistic measures the difference in the skewness and kurtosis of a series to that of the normal distribution (IHS EViews, 2009a). If the residuals are normally distributed, the J-B statistic should not be significant, or the p value should be greater than 0.05 at the 5% level.

Testing for serial correlation

The Durbin-Watson (DW) test is a test for first-order serial correlation where the DW statistic measures the linear association between adjacent residuals from a regression model (IHS EViews, 2009b). The null hypothesis is that there is no serial correlation, while the alternative hypothesis is that there is a first-order serial correlation. The DW statistic ranges from 0 to 4. As a rule of thumb, the residuals are not correlated if the DW statistic is approximately 2, and an acceptable range is 1.50 to 2.50. With 50 or more observations and only a few independent variables, a DW statistic below about 1.5 is a strong indication of a positive first-order serial correlation (IHS EViews, 2009b).

There are a few limitations of the DW test for serial correlation, one of which is that if there are lagged dependent variables included as independent variables, such a test becomes invalid. To overcome these limitations, the Breusch-Godfrey serial correlation Lagrange multiplier (LM) test, which can be used to test for higher order ARMA errors, can be performed. The null hypothesis of the LM test is that there is no serial correlation up to lag order p, where p is a prespecified integer (IHS EViews, 2009b).

Testing for heteroscedasticity

The Breusch-Pagan-Godfrey test is a Lagrange multiplier test of homoscedasticity. The null hypothesis is that the variance of the residuals is the same for all values of the independent variable, while the alternative hypothesis is that the variance of the residuals is different for some values of the independent variable. Accepting the null indicates that the assumption of homoscedasticity is satisfied.

White's heteroscedasticity test is used to test the null hypothesis of no heteroscedasticity against the heteroscedasticity of the unknown, general form (IHS EViews, 2009b). The test statistic is calculated by regressing the squared residuals on all possible (i.e. nonredundant) cross-products of the regressors in an auxiliary regression.

The ARCH test is a Lagrange multiplier test for autoregressive conditional heteroscedasticity (ARCH) in the residuals (IHS EViews, 2009b, p. 162). The null hypothesis is that there is no ARCH up to order q in the residuals.

Testing for model misspecification

Ramsey's RESET (Regression Specification Error Test) is used to test for model misspecification. RESET is a general test for specification errors such as omitted variables, incorrect functional form, correlations between the independent variables, and disturbance errors which may be caused by measurement errors, and the presence of lagged dependent variables and serially correlated disturbances.

3.4.5 Point estimation of long-run elasticities

Elasticity analysis has its theoretical foundation in demand theory and interprets tourism demand from the economic perspective. Such analysis is often carried out to directly benefit policy and decision making. Elasticity measures the responsiveness of tourism demand (i.e. visitor arrivals) from the origins resulting from a change in one determinant. The income elasticity indicates the responsiveness of the tourism demand to the change in the income levels in the origin country or region (Song, Witt, & Li, 2009). Price elasticity has a direct impact on the total revenue, thus it is critical for the suppliers of tourism products and services. When the tourism product is price elastic (i.e. the absolute value being greater than one), the total tourism revenue (TTR, i.e. a product of the average price of the tourism products/services (*P*) and total quantity demanded (*Q*)) increases with a decrease in price. Retrospectively, TTR will decrease when price reduces if the tourism product is price inelastic (i.e. the absolute value being less than one). This is because the percentage change in quantity is less than the percentage decrease in price.

Once the long-run relationship is established, the tourism demand elasticities can be obtained from Equation (3.5):

$$\theta \equiv (\beta_Y, \beta_P, \beta_{PS}) = (-\pi_2 / \pi_1, -\pi_3 / \pi_1, -\pi_4 / \pi_1) \qquad (3.6)$$

where β_{Y}, β_{P}, and β_{PS} represent the income, own price, and cross price elasticities, respectively.

3.4.6 Forecasts of independent variables

Once the model in the current study had passed all the diagnostic tests as specified and all of the coefficients in the models had correct signs, the final step was to calculate the demand elasticities and to forecast visitor arrivals from each source market. Before generating the forecasts of visitor arrivals, the future values of the independent variables including the income, own price, and substitute price variables needed to be predicted first. Forecasts of the real GDP changes published by IMF (2011), as shown in Table 3.2, were used as the projections of the income variable from 2012 to 2015. All values were expressed in real terms. The state space approach to exponential smoothing (SSES) described by Hyndman, Koehler, Ord, and Snyder (2008) were employed to generate the forecasts of the own-price and substitute-price variables. These forecasts of the explanatory variables were then used in conjunction with the estimated relationships to generate the forecasts of the dependent variable (i.e. visitor arrivals) for each source market.

The SSES approach is an innovative framework for automatic forecasting based on an extended range of exponential smoothing methods. This approach has several advantages over traditional exponential smoothing alternatives: (a) it is easy to calculate the model selection criteria such as the likelihood and the AIC within this framework; (b) it provides forecast intervals; and (c) it allows for simulations from the underlying state space model (Hyndman, Koehler, Snyder, & Grose, 2002). Given the merits of this approach, the SSES framework was employed in this study for time series forecasting.

It should be noted that, apart from the model itself, the accuracy of the forecasts may have been subject to the precision of both the GDP forecasts made by IMF and the price forecasts produced by using the exponential smoothing method. Since the variables in the demand models were in logarithm, the forecasting values of visitor arrivals had to be transformed back to natural numbers through the antilogarithm computation. The forecasts of visitor arrivals from all source markets are presented and analysed in Chapter 5.

Table 3.2 Projections of GDP, own price, and substitute price 2011Q1−2015Q4

Region	**Australia**			**China**			**Japan**			**Taiwan**			**UK**			**USA**		
Quarter	GDP	Pi	Ps	GDP	Pi	Ps	GDP	Pi	Ps	GDP	Pi	Ps	GDP	Pi	Ps	GDP	Pi	Ps
2011Q1	3.00	11.03	10.50	9.60	-5.23	8.63	1.40	-7.47	5.36	5.40	-4.07	10.29	1.70	16.29	10.68	2.80	0.52	8.57
2011Q2	3.00	15.06	11.46	9.60	-6.29	12.87	1.40	-8.68	6.24	5.40	-5.05	9.54	1.70	30.99	11.26	2.80	-0.49	10.43
2011Q3	3.00	15.66	9.54	9.60	-4.41	11.48	1.40	-2.31	4.78	5.40	-4.19	5.74	1.70	36.15	9.23	2.80	0.07	8.74
2011Q4	3.00	4.71	6.39	9.60	-4.77	7.85	1.40	-3.14	0.28	5.40	-2.75	1.69	1.70	40.20	5.71	2.80	0.19	5.70
2012Q1	3.50	2.26	5.72	9.50	-4.35	4.32	2.10	-2.85	0.27	5.20	-2.55	2.12	2.30	33.31	4.98	2.90	0.18	5.02
2012Q2	3.50	1.71	5.12	9.50	-3.95	3.72	2.10	-2.58	0.27	5.20	-2.36	2.08	2.30	31.07	4.34	2.90	0.18	4.43
2012Q3	3.50	1.47	4.59	9.50	-3.59	3.33	2.10	-2.34	0.26	5.20	-2.18	2.04	2.30	29.53	3.78	2.90	0.17	3.90
2012Q4	3.50	1.36	4.11	9.50	-3.26	2.95	2.10	-2.12	0.25	5.20	-2.02	2.00	2.30	28.32	3.30	2.90	0.17	3.44
2013Q1	3.30	1.31	3.68	9.50	-2.96	2.75	1.70	-1.92	0.24	5.10	-1.87	1.96	2.50	27.28	2.88	2.70	0.17	3.04
2013Q2	3.30	1.29	3.30	9.50	-2.69	2.58	1.70	-1.74	0.24	5.10	-1.73	1.92	2.50	26.36	2.52	2.70	0.16	2.68
2013Q3	3.30	1.30	2.96	9.50	-2.44	2.45	1.70	-1.57	0.23	5.10	-1.60	1.88	2.50	25.52	2.20	2.70	0.16	2.36
2013Q4	3.30	1.33	2.65	9.50	-2.21	2.29	1.70	-1.42	0.22	5.10	-1.48	1.84	2.50	24.74	1.92	2.70	0.16	2.09
2014Q1	3.30	1.36	2.38	9.50	-2.01	2.19	1.50	-1.29	0.22	5.00	-1.37	1.80	2.50	24.02	1.68	2.70	0.15	1.84
2014Q2	3.30	1.39	2.13	9.50	-1.82	2.11	1.50	-1.17	0.21	5.00	-1.27	1.77	2.50	23.34	1.47	2.70	0.15	1.63
2014Q3	3.30	1.43	1.91	9.50	-1.65	2.03	1.50	-1.06	0.21	5.00	-1.17	1.73	2.50	22.69	1.28	2.70	0.15	1.44
2014Q4	3.30	1.47	1.72	9.50	-1.50	1.94	1.50	-0.96	0.20	5.00	-1.09	1.70	2.50	22.07	1.12	2.70	0.14	1.27
2015Q1	3.20	1.51	1.54	9.50	-1.36	1.88	1.30	-0.87	0.19	4.90	-1.00	1.66	2.60	21.48	0.98	2.70	0.14	1.12
2015Q2	3.20	1.55	1.38	9.50	-1.23	1.82	1.30	-0.78	0.19	4.90	-0.93	1.63	2.60	20.92	0.85	2.70	0.14	0.99
2015Q3	3.20	1.59	1.24	9.50	-1.12	1.77	1.30	-0.71	0.18	4.90	-0.86	1.60	2.60	20.38	0.75	2.70	0.13	0.87
2015Q4	3.20	1.63	1.11	9.50	-1.01	1.71	1.30	-0.64	0.18	4.90	-0.79	1.56	2.60	19.86	0.65	2.70	0.13	0.77

3.4.7 Forecast accuracy of past forecasting exercises

A report published by the HKTDFS evaluated the accuracy of the arrivals forecasts generated from the econometric approach described in Section 3.4. In comparison with the real arrivals figures published by HKTB from 2010Q4 to 2011Q2, Table 3.3 shows that the forecasts published in Volume 2, No. 3 of the HKTDFS forecasting report series were highly accurate as the average MAPE (5.21%) and RMSPE (6.40%) were far below 10 per cent (HKTDFS, 2011). Among the 14 source markets concerned, nine were reported to be less than 10% while three were even less than 5 per cent as measured by MAPE. The largest forecast errors were detected in the case of the Philippines, followed by South Korea, Thailand, Malaysia, and Indonesia. However, none of the MAPEs and RMSPEs exceeded 20 per cent, which again suggests the very good forecasting performance of the ARDL-ECM.

Table 3.3 Accuracy of visitor arrivals forecasts over 2010Q4−2011Q2

Country/Region	MAPE (%)	RMSPE (%)
Australia	2.34	2.49
Taiwan	3.83	4.65
USA	4.09	4.43
UK	6.05	7.28
Singapore	6.32	6.69
Japan	6.64	9.43
China	7.52	9.47
Macau	8.44	9.32
India	9.67	11.31
Indonesia	10.37	11.49
Malaysia	11.24	12.73
Thailand	11.26	12.51
South Korea	11.79	13.96
Philippines	16.96	18.36
Total arrivals in Hong Kong	**5.21**	**6.40**
Mean	7.61	8.78

3.5 Judgmental Forecasting and Adjustments

In Mathews and Diamantopoulos's (1986) study, the inclusion of qualitative inputs in forecasting practice was categorized into three forms. The first form is a *priori incorporation*. This involves the application of judgment before making the forecasts using a quantitative forecasting model, which is manifested in the initial selection of the forecasting model and in the selection, specification, and modification of model parameters. The second form is *concurrent incorporation*, which involves the integration of judgmental and quantitative forecasts using a combining algorithm (e.g. simple averages, weighted averages, or Bayesian analysis). The third form is *a posteriori incorporation*, which involves the revision of a forecast produced by a quantitative model; decision-makers and forecasters modify the forecast results according to their market knowledge to obtain more realistic predictions. The focus of the present study was on the *a posteriori* revision of visitor arrivals forecasts made by econometric models. Specifically, 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 (2000a), is the process of supplying judgmental forecasters with statistical forecasts that they can ignore, accept, or adjust. In this study, the HKTDFS was applied to produce the voluntary integration of statistical forecasts and Delphi experts' judgmental inputs. There were several reasons why the HKTDFS was chosen to effect the integration.

First, the most straightforward reason for choosing the HKTDFS is that it is the first and also the only tourism demand forecasting system available in the Hong Kong tourism industry. Jointly developed by the Public Policy Research Institute and the School of Hotel and Tourism Management at The Hong Kong Polytechnic University, the HKTDFS is a sophisticated Web-based forecasting system aimed at helping the industry achieve a sustainable increase in demand for Hong Kong tourism. It is an innovative system for cooperative and real-time tourism demand forecasting with higher accuracy. The system was officially launched on March 4, 2008 to forecast demand for Hong Kong tourism over the next decade. Since December 15, 2011, the HKTDFS has allowed all registered users (free online registration) to enjoy full permission rights for three functional modules (*Tourism Forecasts*, *Scenario Analysis*, and *Forecasting Adjustment*) and free access to all forecasting reports released. By October 2012, more than 290 users had registered with the system and 110 users had registered since the new policy of user access rights became effective in December 2011. In addition, by November 2012, more than 102,500 visitors had browsed the system.

Second, the econometric (or statistical) methods included in the HKTDFS have already been tested in a few studies, including Chon et al. (2010), Song and Lin (2010), Song et al. (2011), and Song, Lin, and Gao (2012). The findings from these studies show the HKTDFS's ability to handle a variety of data features, obtain reliable models, and produce accurate forecasts for the Hong Kong tourism industry. In addition, parameters in the estimated models and forecasts are updated on a regular base according to the most up-to-date time-series data; for example, arrivals forecasts are updated every year for all 14 selected source markets. The forecasting team has the major role in the model selection, design, and maintenance of the HKTDFS, in particular the algorithms and interfaces, ensuring that the HKTDFS is fit for purpose and can be used effectively in the forecasting process. Forecast accuracy is also evaluated and recorded annually, and the results are published in the

HKTDFS newsletters (e.g. HKTDFS, 2011).

Third, the HKTDFS is easy to use, easy to understand, and easy to improve and extend. The system avoids the display of complicated technical information. Judgmental adjustments in the HKTDFS, where appropriate, are easy to implement and record. Users at various locations can access the system and make real-time adjustments to the forecasts using the built-in statistical tools. The system provides flexible options for adjustments; for example, the system provides an option to make quarterly adjustments to the baseline forecasts, which allows seasonality to be taken into account when the time series is recorded at quarterly intervals and adjustments to be made to cope with more complex seasonality.

Last but not least, since its launch in 2008, the HKTDFS has successfully established and maintained partnerships with a variety of industry stakeholders, including government offices responsible for tourism policy making and implementation; business executives in the travel, hotel, catering, and retail sectors; consultancy firms focusing on the tourism sector; and education and research institutions for tourism. This has helped the HKTDFS to gain a reputation in the Hong Kong tourism industry, which adds to the plausibility of conducting forecasting adjustment via the HKTDFS.

The remainder of this section introduces details about the functional features of the HKTDFS, with a particular focus on the Forecasting Adjustment module. The subsequent section provides the justifications for selecting the Delphi forecasting approach; this is followed by a discussion on the procedure for conducting Delphi forecasting approach in the HKTDFS.

3.5.1 HKTDFS[7]

(1) Overview

The features added to the HKTDFS originally developed by Song, Witt, and Zhang (2008) include the following:

- User-friendliness: the new system architecture can help users to generate tourism forecasts in a more efficient and effective way.
- Modularity: the components of the system are designed as stand-alone modules to reduce the cost of system maintenance.
- Flexibility: the system modules, particularly the application modules, can now be updated and redesigned easily when new technologies and algorithms become available.
- Enhanced website administration system: system administrators and authorized users can log on to the administration module via a Web-based interface and perform routine administrative tasks, such as user account management, file sharing and database management.
- Java Server Pages (JSP) and R-based applications: the JSP web language is used to develop the system, as it provides stable interfaces with external software and languages, such as the statistical language and the R environment.
- Implementation of open source R code: R provides a wide variety of statistical and graphical options, including linear and nonlinear modelling, classic statistical tests, time series analysis, classification, clustering, and many other statistical applications. With HKTDFS, the quantitative forecasts are generated in the R environment.

[7] Parts of this section was published in Song, Gao, and Lin (2013, pp. 297-299 & 301-302).

- Improvements in judgmental inputs: the system includes a dynamic online Delphi survey module, which allows the integration of statistical and judgmental forecasts.

Figure 3.3 illustrates the four-tier Client/Server (C/S) architecture of the HKTDFS Web platform. The first two tiers are traditional components of the C/S architecture, by which users interact with the system for any specific application. In the third (business) tier, the core functions, including all operational logistics, are hosted on an Apache Tomcat Web Server. In particular, an interface known as REngine, an abstract base class for all implementations of R engines, is deployed in this tier to allow communication between the web platform and the R environment. To begin with, a dataset (in Excel format) is supplied by the user. Once a request to estimate the model is given, the system connects itself to the REngine client and runs the Model Estimation module. The estimation results, including the diagnostic tests and tourism demand elasticities (e.g. income and price elasticities), are then presented on the web pages. The results are stored in the database simultaneously (see Figure 3.3).

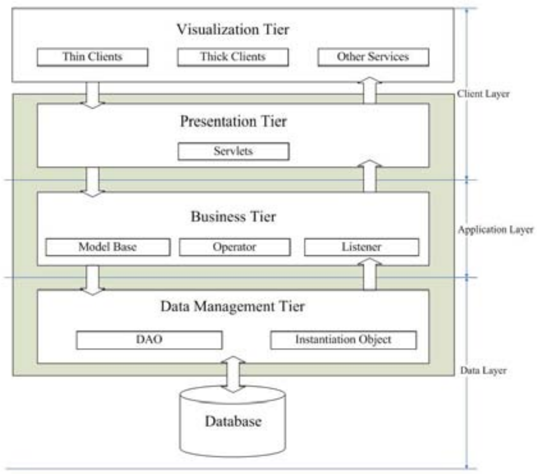


Figure 3.3 HKTDFS architecture

The fourth tier is the database tier (see Figure 3.4). Two different databases are available in this tier. The first is the MySQL database, which is used to store three types of data: (a) historical time series of all tourism demand measures and their influencing factors, such as GDP, own price and substitute prices; (b) estimation results (e.g. diagnostic statistics and elasticities); and (c) forecasts. The second database contains the R program codes that are used to run the econometric model (or ARDL-ECM) embedded in the system.

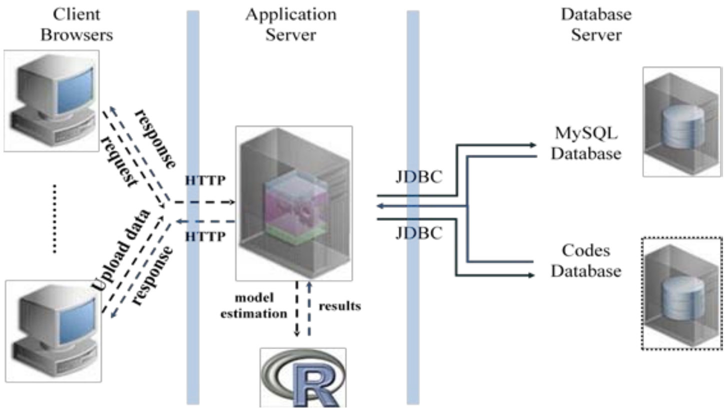


Figure 3.4 HKTDFS flowchart

Although most tourism managers/forecasters have rich industry experience, some may have very little knowledge about quantitative forecasting methods, and particularly about advanced econometric forecasting methods. To make the use of the HKTDFS easier, the system is designed to automate the forecasting process, and thus requires little modelling knowledge. The system also makes full use of the forecasters' domain knowledge and integrates it with the econometric forecasts in order to achieve greater forecast accuracy.

The forecasting procedure in the HKTDFS involves three stages (see Figure 3.5). The first stage is the pre-modelling data analysis, which is performed outside the HKTDFS via a number of statistical analysis/software packages such as SPSS, EViews and Excel. The tasks in this stage are to examine and identify the properties of the data by testing for unit roots and co-integration, and to introduce dummy variables that take seasonality and the impact of special events into account. Following the preliminary data analysis, the processed dataset can then be imported into the HKTDFS.

In the second stage, once the data have been input into the system and are ready for the forecasting tasks to be performed, the system runs the Model Estimation module automatically after receiving the user's HTML instructions. The user can select the model by choosing different dependent and independent variables and conducting diagnostic checks on the model adequacy. The general-to-specific methodology is followed to obtain the final ARDL-ECM model, which passes most of the diagnostic tests. After the final model has been confirmed by the user, forecasts of the dependent variables are generated and stored in the system database.

In the third stage, users can adjust the statistical forecasts based on their domain knowledge if they believe that there is important information which is not captured by the econometric model. The HKTDFS consists of three functional modules: the data module, the quantitative forecasting module, and the judgmental forecasting module (see Figure 3.5). Detailed descriptions of data module and quantitative modules can be found in Song, Gao, and Lin (2013). This study only focused on the judgmental forecasting module.

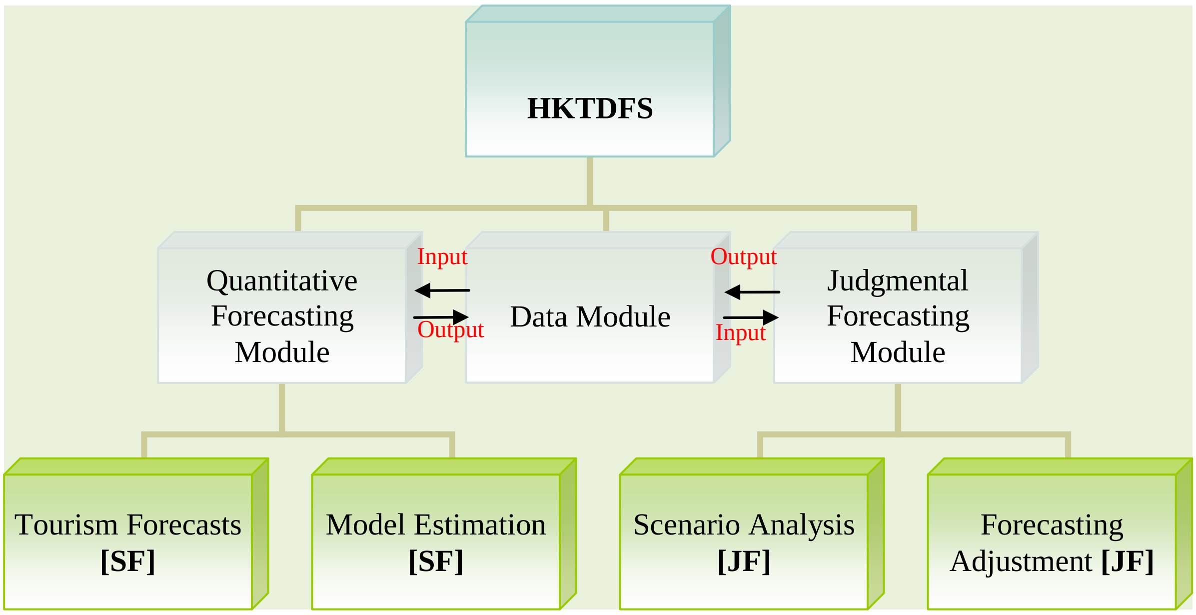
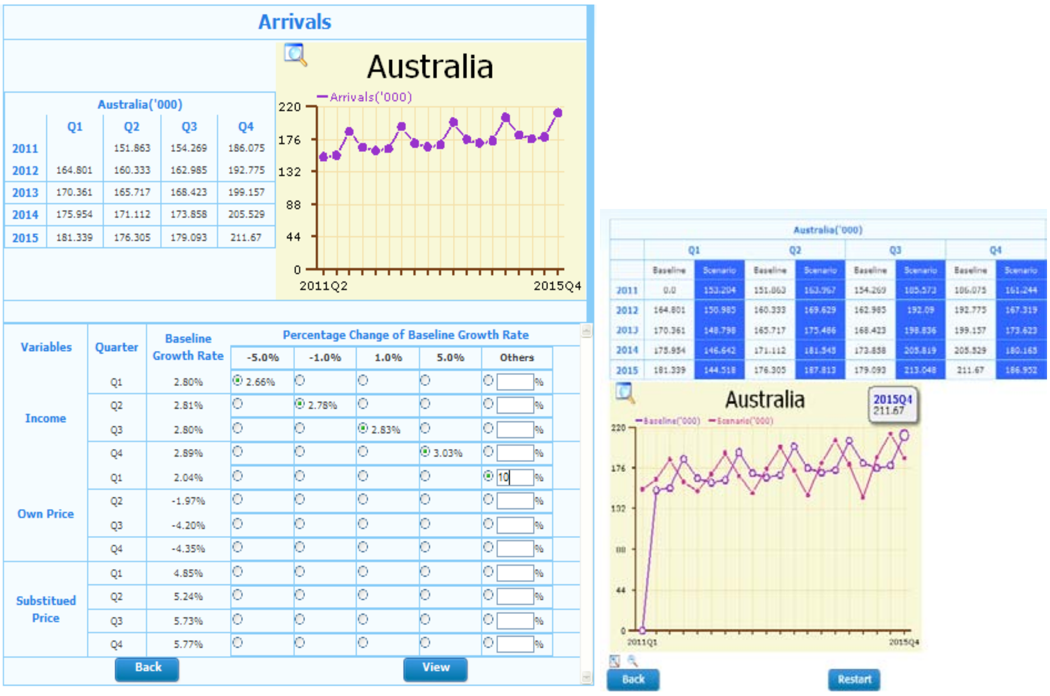


Figure 3.5 HKTDFS components

(2) Judgmental forecasting module

After the statistical forecasts have been produced, the system allows users to incorporate their domain knowledge in them. Two modules, *Scenario Analysis* and *Statistical Adjustment*, are available to users for entering their judgmental inputs into the system. The *Scenario Analysis* module takes the statistical forecasts provided by the ARDL-ECM as the baseline forecasts, and these forecasts are then used as benchmarks for the scenario forecasts created by the user. This component offers four baseline scenarios (5% or 1% higher or lower than the benchmark growth rates), plus a customized scenario where users can input their own estimates (see Figure 3.6 (a)). When a specific scenario is submitted, the system will present the scenario forecasting results and the baseline statistical forecasts (see Figure 3.6 (b)) in both tabular and graphic formats. The system also allows users to revise the statistical forecasts by going back to the Model Estimation module.



(a) (b)

Figure 3.6 Screen shots of the scenario analysis

Unlike the Scenario Analysis module, the Forecasting Adjustment module allows users to adjust the forecasts of both the dependent and independent variables. This component is also responsible for a dynamic Delphi survey procedure and includes a few features that are not included in the earlier HKTDFS version described by Song, Witt, and Zhang (2008). These features include the following.

First, the system presents both the historical data series and the forecasts, to permit experts to compare the historical trends of the time series with the forecasts easily. According to Benson and Önkal (1992) and Fildes, Goodwin, and Lawrence (2006), giving experts access to the latest observations of the time series can improve the accuracy of the adjusted forecasts.

Second, the system allows the experts to give the reasons for their adjustments to the statistical forecasts. Previous studies have suggested that recording these reasons is an effective way to structure the adjustments and improve the accuracy of judgmental forecasts (Armstrong, 2006; Goodwin, 2000b; Rowe & Wright, 1999).

Third, the system also allows group feedback to be recorded, permitting the experts to refer to it during the later rounds of forecasting adjustments. O'Connor (1989), Lawrence et al. (2006), and Rowe and Wright (1999) concluded that feedback improves the accuracy of statistical forecasts.

As suggested by the DSS literature, two broad approaches, namely *restrictiveness* and *decisional guidance* can be used to design an FSS in order to achieve the key objectives of improving the forecaster's ability to realize when judgmental intervention is appropriate and enabling system users to apply accurate judgmental inputs when appropriate (Fildes, Goodwin, & Lawrence, 2006).

Silver (1991) defined system restrictiveness as the way a DSS "limits its users' decision-making processes" whereas decisional guidance as how a DSS "guides its users in structuring and executing their decision-making processes by assisting them in choosing and using the system's functional capabilities" (p. 108). Restrictiveness can determine the manner in which forecasts are obtained by limiting or denying the user the opportunity to employ particular processes or requiring that alternative processes are adopted (Fildes, Goodwin, & Lawrence, 2006). In general, the more restrictive an FSS, the less the opportunity for providing guidance. As a consequence, it requires a design trade-off regarding the interaction between restrictiveness and guidance that for each judgmental opportunity, the designer must decide whether to restrict the decision-making process or provide guidance on it (Silver, 1991). The concepts of restrictiveness and guidance are applied in the HKTDFS design.

As suggested by Fildes, Goodwin, and Lawrence (2006), an ideally designed will have the following five attributes: acceptable to users, easy to use, a flexible range of appropriate forecasting methods and facilities, viable for commercial software companies to market, and foster the appropriate mix of judgmental and statistical methods. Fildes, Goodwin, and Lawrence (2006) argued that restrictiveness would be unlikely to lead to an FSS that allows for the coexistence of these attributes. For example, one FSS may be easy to use but may fall far short of the highly flexible and multifaceted support system.

In Forecasting Adjustment module, the statistical forecasts generated by the ARDL-ECM approach are provided as the baseline system forecasts. Users are not allowed to select other forecasting method. Past studies suggest that when users have the ability to choose the statistical model with which to produce their forecasts, it is often their choice is quite poor (Fildes, Goodwin, & Lawrence, 2006). It is further noted by Fildes, Goodwin, and Lawrence that forecasters often select the default parameter values or sub-optimal methods, and they attempt to make large judgmental adjustments to the quantitative forecasts even they may probably be unnecessary.

Only Delphi experts, authorized users and full subscribers are able to access the Forecasting Adjustment module. In each round of the Delphi survey, this module provides panellists with two alternative ways of making their judgmental adjustments: (a) by changing the point forecasts of the dependent variables by year or by individual quarters over the specified forecasting period, and (b) by changing the growth rates of the determinant variables of tourism demand, as in the Scenario Analysis module. Upon the completion of each round of the Delphi survey, the module publishes the final group forecasts (or median forecasts), which can be accessed by all of the experts.

3.5.2 Delphi method

(1) Overview

As discussed earlier in Chapter 2, the use of the structured group technique is likely to yield more accurate forecasts than the use of the unstructured approach. When a group of people has to decide what adjustments to apply to a set of statistical forecasts, one way of avoiding the biases that occur in meetings is to use the Delphi method (Goodwin, 2005). Therefore, in this study, the Delphi technique was used to produce judgmental forecasts.

By administering a series of questionnaires, the Delphi technique aims at combining the knowledge and experience of a selected group of experts in the areas of interest to form a consensus of opinion about the likely occurrence of specified time periods. A panel of experts was selected from different stakeholders of the tourism industry in Hong Kong, including two groups, namely industry practitioners and academic researchers who had knowledge and experience in forecasting and also worked closely with forecasting tasks. The advantages of using the Delphi approach to generate judgmental forecasts are: respondent anonymity (reducing the dominant members' effect), and iteration and controlled feedback from the respondents (Frechtling, 2001).

This study adopted the definitions of contextual knowledge and technical knowledge from Sanders and Ritzman's (1992) study. *Industry experts* were defined as those industry practitioners with relatively less technical knowledge but more contextual knowledge (i.e. general forecasting experience in the industry and product knowledge about the specific items involved) gained by performing the same forecasting function as part of their job; thus, they are more sensitive to a variety of cause-effect relationships, environmental cues, and other organizational information, all of which might affect the variable being forecast.

Academic experts were defined as those academic researchers with relatively richer technical knowledge but less contextual knowledge. They know more about data analysis and formal forecasting procedures, including information on the logic, capability and use of the various statistical techniques that can be applied to time-series data as well as information on how to analyse data judgmentally. Knowing about the different components of demand, outliers, autocorrelation, and biases inherent in human judgment is also part of this technical knowledge.

The most important element of the Delphi approach is to record experts' forecasts, which are then distributed in the subsequent round for panel members' reconsideration. This process normally continues for two to four rounds until a consensus is reached among all of the experts. In the initial round of forecasting, it is likely that the forecasts given by the experts will be widely distributed. In successive rounds, the distribution of responses will converge towards the mean or median values.

(2) Procedure

A two-stage research framework presented in Figure 3.7 was applied to conduct Delphi forecasting in this study. In **Stage I**, the first step was to identify the problem statement. A background paper that briefly introduced the series of interest was incorporated into the first-round questionnaire and circulated to the panellists. The second and third steps were to determine the make-up of the expert panel and the sample size of the panel.

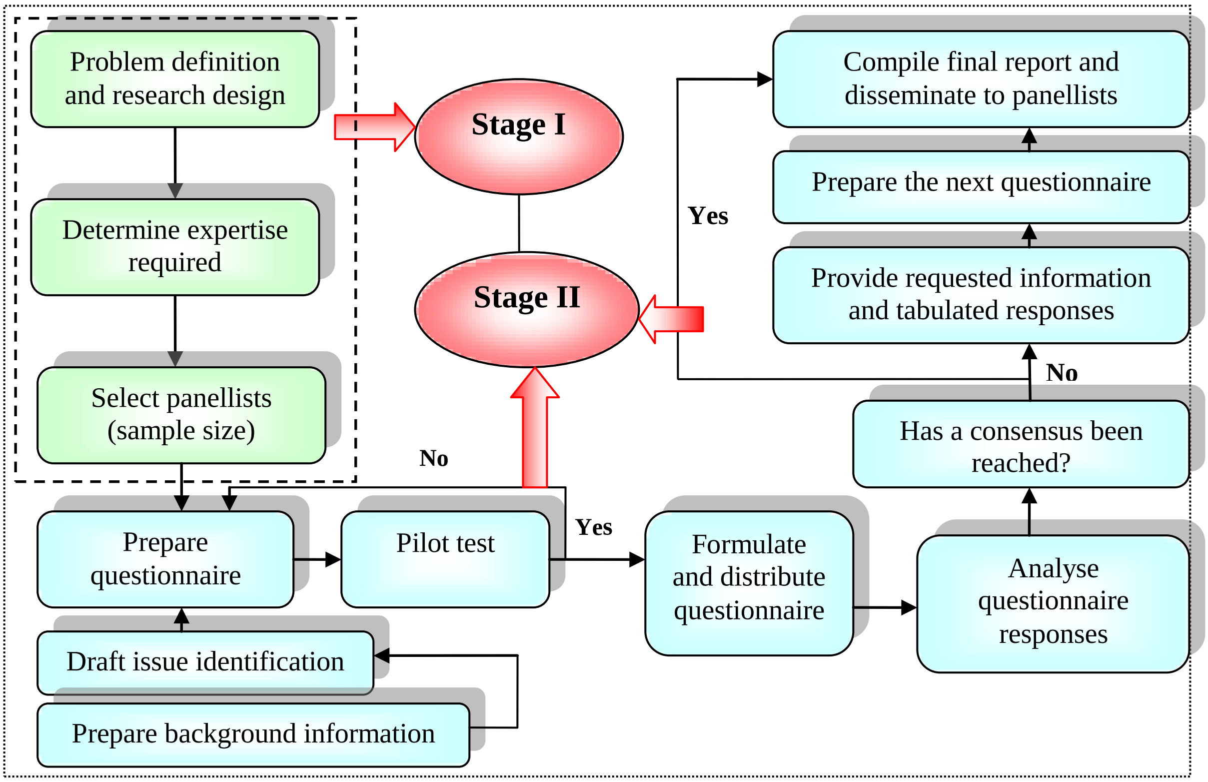


Figure 3.7 Delphi research design

Source: Adapted from Kaynak and Macaulay (1984, p. 94).

Table 3.4 provides detailed steps on how to select the Delphi members. The initial formulation of the Delphi panel list featured a review of the relevant literature and an extensive search of the directory information from the official website of HKTB (http://partnernet.hktb.com). A group of heterogeneous experts from academic institutions, public and private sectors of tourism industry in Hong Kong, whose combined knowledge and expertise reflecting the full scope of the tourism forecasting domain, were selected. Potential panel members were reached by email and phone to seek their acceptance. Follow-up letters along with the instructions on how to conduct Delphi surveys were emailed to those experts who agreed to participate.

Table 3.4 Procedure for selecting Delphi panel members

Step	Procedure	Result
Step 1	Identify potential experts: • Review literature to compile a list of potential panel members from academic researchers based on their recent since 2000 (mainly books and journal articles) on modelling and forecasting Hong Kong tourism demand. • Search researchers whose research areas included tourism economics, tourism forecasting, and relevant topics from the official websites of four universities in Hong Kong: The Hong Kong Polytechnic University, The University of Hong Kong, The Chinese University of Hong Kong, and City University of Hong Kong. • Check the official website of HKTB for a comprehensive directory with detailed information on all business partners in the Hong Kong tourism industry.	List of names.
Step 2	Send email invitations to potential panellists identified from **Step 1**. Distribute stamped, self-addressed envelopes to the local experts to ensure a higher return rate. Briefly explain purpose, scope, and significance of the project in the invitation letter.	Develop a final list of potential panellists to participate in study and a list of substitutes. A total of 32 respondents (17 industry people and 15 academic researchers) agreed to take part in the survey.
Step 3	Conduct a pilot study among postgraduate students and research staff to test the reliability of the Statistical Adjustment module in the HKTDFS. Carry out feedback survey to examine respondents' perceptions of the effectiveness of using the HKTDFS to conduct forecasting tasks in tourism.	Pilot study survey involving 21 students and 5 research staff.
Step 4	Send instructions to the panellists in the final list obtained from **Step 2**. Invite those panellists to conduct the main Delphi survey.	

Before conducting the main Delphi survey, a pilot study was conducted among postgraduate students in the School of Hotel and Tourism Management at the Hong Kong Polytechnic University. The purpose of this pilot testing exercise was to examine the reliability of the HKTDFS's Statistical Adjustment module that would be used in the main Delphi survey. To recruit participants, letters were sent to a group of postgraduate students enrolled in a doctoral-level quantitative methods course and to researchers working in tourism-related projects. The final panel consisted of 21 students and 5 research staff members.

Six source markets were selected, and it was decided that the forecasting period would be five years over 2011Q1–2015Q4 because this was felt to be a period from which respondents could identify the current trend in terms of magnitude of change. After making the necessary revisions in the HKTDFS, the instructions were emailed to the participants to start the survey.

Panellists were invited to make their adjustments to the econometric forecasts of visitor arrivals from three short-haul markets (i.e. China, Taiwan, and Japan) and three long-haul markets (i.e. the USA, the UK, and Australia) of Hong Kong. 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. All these events were listed at the end of the instructions (see Appendix A).

In **Stage II**, the questionnaire was compiled and adapted to obtain all forecasting series in the HKTDFS. The questioning process was similar for both rounds of the Delphi survey. The first round of questions, which included a description of the intentions and purpose of the study, was the critical stage of the survey. A follow-up letter was sent as a reminder to those who did not respond three days before the deadline for each round. Participants were required to submit their forecasts (adjustments) privately and independently to the HKTDFS. An assessment was then made as to whether there was a consensus, in which case the median estimate was used to summarize group experts' judgmental adjustments. Descriptive statistics, including the mean, median, maximum and minimum values, were also used to summarize the group's forecasts. The summarized results were then distributed to the participants for their reconsideration in the subsequent round. The process was repeated until the participants reached a consensus or few people were changing their forecasts. The final stage of the survey was to verify and generalize the research results, compile the final report and disseminate it to all of the panellists.

(3) Profile of the Delphi panellists

Industry professionals and academic researchers were contacted by email and informed about the study's objectives. By March 2011, 32 of the 950 people contacted by email had agreed to serve on the panel. These 32 experts came 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 (see Table 3.5). Some of the contacts could not be reached after repeated attempts, several felt that they did not have time to participate, and others declined to participate because of a perceived lack of expertise. Fourteen participants dropped out during the course of the first round due to schedule clashes or other personal reasons.

Table 3.5 Composition of the Delphi panel

Sector	**Initial contact**	**R1**	**R2**
Academic institutions	15	11	11
Accommodation	6	2	1
Government	2	2	2
Tourist attractions/facilities	4	2	2
Travel trades	5	1	1
Grand Total	**32**	**18**	**17**

3.6 Research Hypotheses

This study adopted a standard approach to analysing forecasting performance—investigate the bias of forecasts and the efficiency of forecasts in terms of incorporating all available information and assess their forecast accuracy. The traditional terminology of forecast evaluation is to evaluate forecast accuracy (or forecast errors). In many contexts, accuracy is the top concern in forecasting performance; however, measurements of accuracy do not offer guidance on how to improve forecasts (Musso & Phillips, 2002). Tests for bias are intended to check whether forecasts tend to lean one way or the other. Tests for efficiency are intended to check whether forecasts have taken all available information into account; if forecasts are efficient, there should be "no correlation between any variable measured when the projections are formed and the error later observed" (Musso & Phillips, 2002, p. 24). To achieve the research objectives outlined in Section 1.3 of Chapter 1, a number of research hypotheses were developed, and these are described below.

3.6.1 Hypotheses on the accuracy of judgmentally adjusted forecasts

One major challenge for the designers of forecasting support system in tourism is to combine the stability and consistency of statistical forecasting techniques with the need for expert judgment. The review presented in Section 2.3 shows that the integration of statistical and judgmental forecasting methods is likely to improve accuracy significantly over individual forecasts (Armstrong, 2001a; Blattberg & Hoch, 1990; Lawrence et al., 2006; Lim & O'Connor, 1995; Lobo & Nair, 1990). The integration process can be in the form of either a mechanical combination or an adjustment of statistical forecasts (Lim & O'Connor, 1995). The latter mode was adopted in the current study.

Sanders and Ritzman (2001) 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 ad hoc. An 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 findings are presented in Chapter 4. Based on their findings, hypothesis *H1a* was formulated:

H1a: Judgmental forecast adjustments based on statistical forecasts improve forecast accuracy.

The relative accuracy of statistical forecasts compared to those generated by the simplest Naive model 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 Naive no change model. Thus, the Naive 1 model was used as a basis of comparison for forecasting evaluation in this study and hypothesis *H1b* was developed:

H1b: Judgmentally adjusted forecasts are more accurate than Naive forecasts.

3.6.2 Hypotheses on the bias and inefficiency of judgmentally adjusted forecasts

The term *judgmental heuristic* is defined as a strategy that "relies on a natural assessment to produce an estimation or a prediction" (Tversky & Kahneman, 2002,

p. 20). Tversky and Kahneman (1974) originally examined three main heuristics – availability, representativeness, and anchoring-and-adjustment – commonly used in probability assessments and value prediction. Tversky and Kahneman (2002) showed that the use of judgmental heuristics gives rise to predictable biases.

Before using human judgment to improve forecast accuracy, it is necessary to understand its biases and limitations along with its major advantages. An examination of two properties, unbiasedness and efficiency, allows for integrating the information from statistical forecasts with experts' judgments by exploiting the advantages of both while avoiding their drawbacks (Clemen, 1989). Lawrence et al. (2000) concluded that two major sources of error were bias and inefficiency (in that there was a serial correlation in the errors) in the forecasts, and these two factors seemed to mask any contribution of contextual information to accuracy. Using tests that determine whether forecasts are unbiased and efficient shows whether it would have been possible to improve upon observed forecast accuracy.

A number of empirical studies have been carried out to examine the effectiveness of judgmental adjustments to statistical forecasts, and the results have been mixed. Lawrence, O'Connor, and Edmundson (2000) found that studies on real world judgmental forecasting all reported bias and inefficiency in the forecasts. Musso and Phillips (2002) stated that unbiasedness is "a necessary, though not sufficient, condition for efficiency" (p. 25), suggesting that efficiency is a more demanding criterion than unbiasedness. Unbiasedness and efficiency 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 et al. (2009) also found that although judgmental adjustments would probably help to improve accuracy, they may be either biased or inefficient. Thus, the following hypotheses were formulated:

H2a: Judgmentally adjusted forecasts are biased.

H2b: Judgmentally adjusted forecasts are inefficient.

3.6.3 Hypotheses about the Delphi process

The theoretical explanation for why the Delphi method can provide accurate forecasts is the "theory of errors" proposed by Dalkey (1975, cited in Rowe, 1998). Based on Dalkey's theory of errors, Parente and Anderson-Parente (1987) theoretically demonstrated how the Delphi technique could improve judgmental accuracy over rounds. Studies in the Delphi forecasting literature have attempted to evaluate Delphi under the assumption that this technique is intended to improve accuracy (Rowe, 1998). Parente and Anderson-Parente (2011) illustrated that the basic assumption in applying Delphi forecasting methods is that consensus forecasts obtained from structured groups will be more accurate than those derived from "at least half of the group" (p. 1705).

The central problem in the variable weighting of experts' judgments is determining how to weigh them. The simplest but most common way of obtaining forecasts from a Delphi procedure is to average the forecasts made by all individual panellists without interaction (Rowe, 1998). Rowe and Wright (2001) suggested that final forecasts should be obtained by weighting all of the experts' estimates equally and aggregating them. This is equivalent to the average of the equally weighted estimates of the members of a statistical group.

Rowe (1998) summarized two benefits of using such a statistical (or “statistized”) group. First, compared to the use of a randomly selected individual, the use of a statistical group can increase the reliability of forecasts. Furthermore, it would possible to average out random errors and produce a response centring upon the true value of the group estimate if an assumption is based on the belief that “a ‘true plus error’ model adequately describes individual estimates for a particular problem” (Rowe, 1998, p. 210), and this may result in a judgment that is better than that of the best individual panellist.

Due to the problem of systematic bias typically observed in human judgmental performance, Rowe (1998) suggested that it is more appropriate to use a “bias plus error” model. By using such a model, averaging individual estimates will not eliminate the mean error (i.e. the average estimate will be centred upon the mean of “erroneous judgment” rather than the true value) but is likely to still result in an improved judgment.

The majority of the Delphi evaluative studies were carried out by comparing the final round aggregate outcome either to derive first round judgments that were equal to the aggregate judgments from noninteracting groups or to aggregate judgments from interacting groups (Rowe, 1998; Woudenberg, 1991). However, the findings from the above comparisons are equivocal. A solid body of research supports the advantage of Delphi groups over traditional groups and statistical groups (e.g. Rowe & Wright, 2001). The value of the Delphi technique is likely to be greater when unexpected information could be provided to the individual experts or be obtained from other members of the group. For example, Rowe (1998) provided such a review in comparing Delphi to statistical groups (i.e. the average of individual estimates) and interacting groups (e.g. NGT) and found that, in general, Delphi can yield significantly higher accuracy than either of the other two techniques. However, there is no consistent evidence to support the argument that Delphi can produce superior accuracy to all other structured group techniques. In Woudenberg's (1991) study, it was found that Delphi was more accurate than unstructured, direct interaction groups, but less accurate than statistical groups. In addition, Woudenberg (1991) also found that there was no difference in accuracy between Delphi and structured, direct interaction groups.

Studies that compared Delphi accuracy over rounds have generally provided evidence that significant increases in accuracy are attained over different rounds (Rowe, 1998; Rowe & Wright, 1999; Woudenberg, 1991). For example, Rowe and Wright (1999) extensively reviewed 21 published studies on Delphi through 1999, and found that more than half of the studies reported a higher accuracy level after iterative polling compared to that derived from the initial round. Among the 21 studies, only two reported the reverse effect. In addition, Delphi accuracy has also been evaluated by comparing the overall group performance to that of the individual panellists who provide the best performance (Rowe, 1998). Studies on this topic have also reported mixed results.

Based on the above findings, the following two hypotheses were formulated:

H3a: Forecast accuracy improves via the Delphi approach: Final Delphi forecasts are more accurate than the average of the initial estimates of the group members (i.e. statistical group).

H3b: Combining judgmental forecasts with the mean is more accurate than other aggregating measures.

3.6.4 Hypotheses on self-rated expertise

As discussed in the previous section, experts' opinions are usually combined through an equal weighting scheme. It is also possible to weigh panellists' estimates differentially or unequally. In real world forecasting, objective measures of expertise are unlikely to be available except when the forecasting task is repetitive and there are detailed records of past performance (e.g. weather forecasts) (Rowe & Wright, 2001). A common method of identifying experts is to rely on their self-rated expertise (i.e. a judgment by the individual of his/her competence or knowledgeability concerning the estimate) (Linstone & Turoff, 2002). However, whether forecasting exercises should be based on "top-expert" assessments or on a broader base of less specialized experts, and whether self-rating is an acceptable method are still controversial issues (Tichy, 2004). Larréché and Moinpour (1983) showed that self-rated expertise may be a more appropriate measure when experts can actually evaluate their expertise in terms of a specific problem area to which they are regularly exposed.

The use of self-rated expertise has turned out to be a significant index for rating group estimates; however, it has not been included in a formal theory of aggregation due to the lack of a theoretical definition of self-rating (Linstone & Turoff, 2002). Rowe and Wright (2001) illustrated that a weighting scheme based on objective data, such as experts' self-ratings of expertise or confidence, were not generally shown to be valid indicators of expertise in judgmental forecasting tasks. However, self-rating has proved to be valuable for selecting more accurate subgroups (e.g. Dalkey, Brown, & Cochran, 1969; Best, 1974; Rowe & Wright, 1996).

In the Delphi forecasting literature, the evidence concerning how to select subgroups based on self-ratings is somewhat inconsistent. Larreché and Moinpour

(1983) demonstrated that one could achieve better accuracy in an estimation task by aggregating only the estimates of those identified as most expert according to an external measure of expertise. Some studies have shown that the Delphi approach can lead to increased accuracy of group responses more often than not and that a self-rating index (the average of individual self-ratings on a given question) is a valid indicator of the mean accuracy of group responses (Dalkey, Brown, & Cochran, 1969). Delphi procedures have also been found to be more effective if self-rating information is used to select more accurate subgroups (Dalkey, Brown, & Cochran, 1969). In a laboratory setting, Best (1974) concluded that self-rated experts made significantly more accurate estimates than self-rated nonexperts for both past demand and student enrolment. However, one must raise the question of whether an experiment based on almanac-type questions serves as an adequate basis from which to draw conclusions about the validity of self-rated expertise in Delphi forecasting (Linstone & Turoff, 2002). While lognormality behaviour exhibits a similar pattern for factual and forecasting cases, this similarity might not carry over for self-ratings.

Based on Rowe and Wright’s (1996) findings, the following two hypotheses were proposed:

H4a: Higher self-rated expertise is related to more accurate Round 1 forecasts.

H4b: Higher self-rated expertise is related to a lower propensity to make judgment changes over rounds.

3.6.5 Hypotheses on the characteristics of judgmental forecasting tasks

Numerous models and theories have been proposed in various fields to gain a complete understanding about the determinants of human behaviour (O’Connor, 1989). Generally, the importance of the individual and the nature of the task in influencing behaviour have been emphasized (O'Connor, 1989). Sanders and Ritzman (1992) assessed the effects of three situational variables − data variability, contextual knowledge, and technical knowledge − on the relative forecasting performance of judgmental forecasts. Diamantopoulos and Mathews (1989) observed that one major factor determining the effectiveness of forecast adjustment is related to product-specific circumstances. They also demonstrated the importance of situational factors (e.g. time horizon, data availability, and product type) in forecasting.

Many studies have agreed that the relative accuracy of judgmental forecasts could be determined by the characteristics of the data series to be forecast (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009; Lim & O'Connor, 1995; Sanders & Ritzman, 1992), for example, the historical stability of a series or the data variability of a time series. Sanders and Ritzman (1992) summarized two sources of the variability in a time series, one coming "from variability in causal factors that affect the dependent variable being measured by the time series" and the other one being "the inherent randomness and uncertainty that cannot be explained by causal factors or by accounting for auto-correlation patterns" (p. 42). Sanders and Ritzman (1992) further illustrated that contextual and technical knowledge could help to explain the first source of variability but not the second one.

Schnaars (1984) stated that forecasts of unstable series are notoriously inaccurate, especially if complex extrapolations are used. Sanders and Ritzman (1992) found that the statistical method performs better than the judgmental forecasts made by practitioners for stable time series; however, judgmental forecasts made by practitioners become more preferable as the variability of a time-series increases. The following two hypotheses were thus proposed to examine the relationship between accuracy and data variability:

H5a: Forecast accuracy decreases with data variability.

H5b: Experts' intervention is more valuable for more variable time series.

In practice, forecasters usually have certain information or knowledge about the forecasting tasks in addition to the time-series values. The additional information, also called "non-time-series" information, was categorized as "contextual" knowledge and "technical" knowledge by Sanders and Ritzman (1992). They defined contextual knowledge as "knowledge that practitioners gain through experience as part of their jobs" and stated that "practitioners become sensitive to a variety of cause-effect relationships, environmental cues, and other organizational information" (Sanders & Ritzman, 1992, p. 40). Technical knowledge is described as involving "data analysis and formal forecasting procedures", including "information on logic, capability, and use of the various statistical techniques that can be applied to time-series data as well as information on how to analyse data judgmentally, such as visual checks for trends, runs, and cyclical behaviour" (Sanders & Ritzman, 1992, p. 40). Knowledge on identifying the different components of demand, outliers, autocorrelation, and the biases inherent in human judgment is also regarded as part of technical knowledge. Edmundson, Lawrence, and O'Connor (1988) divided contextual knowledge into two aspects: product knowledge about the specific items involved and general forecasting experience in the industry pertaining to an understanding of the cause-effect relationships involved; Webby and O'Connor (1996) called the latter "causal knowledge".

The value of the contextual information possessed by forecasters/managers but not by the statistical prediction was explored by Yaniv and Hogarth (1993) in a laboratory study that involved completing word fragments. In their study, as the level of contextual information increased, the accuracy of the judgments of the forecaster relative to the quantitative model also increased.

Without contextual information, judgmental adjustment may reduce the forecast accuracy due to anchoring and adjustment heuristics (Webby & O'Connor, 1996). Some studies showed that even without domain knowledge, forecasters who were able to recognize pattern changes could make judgmental adjustment that led to improvements in the forecast accuracy (Sanders, 1992). However, it is more likely to improve accuracy when adjustments are made based on domain knowledge (Sanders & Ritzman, 2001). Sanders and Ritzman (1992) concluded that judgmental forecasts based on contextual knowledge were significantly more accurate than those based on technical knowledge or no such knowledge, and were even superior to the statistical forecasts. Other researchers have arrived at similar conclusions with regard to the benefits of domain knowledge − examples include Edmundson, Lawrence, and O'Connor (1988), Fildes and Goodwin (2007), Marmier and Cheikhrouhou (2010), and Mathews and Diamantopoulos (1986, 1989, 1990).

In addition to domain knowledge, there are other factors that may affect the accuracy of judgmental adjustments, including feedback (Remus, O'Connor, & Griggs, 1996), incentive (Remus, O'Connor, & Griggs, 1998), excess error (Mathews & Diamantopoulous, 1990; Willemain, 1989, 1991), data presentation form (Harvey & Bolger, 1996), and task structure (Angus-Leppan & Fatseas, 1986).

Some studies have concluded that the integration of judgmental and statistical forecasts is likely to lead to significant accuracy improvement when contextual information is available (Guerard & Beidleman, 1987; Lobo, 1991). Sanders and

Ritzman (1995) showed that judgmental forecasts based on contextual knowledge rather than technical knowledge are considered as better input to the integration process. The above reasoning leads to the following hypothesis on contextual knowledge:

> ***H5c: Judgmentally adjusted forecasts made by experts with more contextual knowledge are more accurate than those made by experts with less contextual knowledge.***

The combination should be particularly useful for long-range forecasts as the uncertainty increases with the forecasting horizon. Makridakis and Winkler (1983) found that reductions in MAPE were decreased as the forecasting horizon increased. Lobo (1992) examined quarterly earnings forecasts for 205 firms over 1978–1985 and found that, for one-quarter-ahead forecasts, the average MAPE for combined forecast was 32.3%, 4.5% smaller than that of the average component forecasts.

Harvey (2007) pointed out that forecasts that are made further ahead in time are more subject to error. Joutz and Stekler (2000) analysed the relationships between the accuracy of USA Federal Reserve forecasts and the length of the forecasting horizon, and they found that forecasts made later in the quarter were more accurate in terms of forecasts for the current quarter but not for subsequent quarters. In examining the effect of a decrease in the forecasting horizon on the informational efficiency of analysts' forecasts of annual earnings per share, Ali, Klein, and Rosenfeld (1992) found that the analysts they studied showed a remarkable improvement in their forecasting ability as the forecasting horizon shrank from 8 months to one. Based on the above empirical evidence, it was hypothesized as follows:

H5d: Improvement in accuracy from judgmental adjustments decreases over time.

3.6.6 Hypotheses about adjustment behaviour

Forecasting scholars have investigated why forecasters decide to apply a judgmental adjustment to particular statistical forecasts. Studies have shown that judgmental forecasters can recognize forecasts that are in need of adjustment even when they only have access to time-series information (Willemain, 1989). Mathews and Diamantopoulous (1990) found that managers are able to select the most inadequate system forecasts and then adjust them in the correct direction. In a more recent study, Fildes et al. (2009) compared unadjusted forecasts with forecasts that were subsequently judgementally adjusted and found that forecasters were able to identify forecasts that were most in need of adjustment. We therefore propose the following hypothesis:

H6a: The forecasts selected for adjustment are those most in need of adjustment.

Diamantopoulos and Mathews (1989) revealed that differences between individuals in terms of personality traits (e.g. optimism/pessimism, attitude to risk, and self-confidence), experience in the industry, familiarity with the products concerned, market knowledge, and education can influence the direction and magnitude of the adjustment undertaken and hence the nature and effectiveness of the forecasting adjustment process. These characteristics determine, to a certain extent, the individual forecaster’s appreciation of the shortcomings of quantitative forecasting models, which will help him/her to identify potentially poor forecasts.

Empirical studies have shown that judgmental forecast accuracy may be affected by judgmental behaviour, such as the size of the adjustment (Diamantopoulos & Mathews, 1989; Fildes et al., 2009; Mathews & Diamantopoulos, 1986, 1990, 1992), and the direction of the adjustment (Fildes et al., 2009). For example, Diamantopoulos and Mathews (1989) found that larger adjustments are more effective in improving accuracy than smaller ones. However, this finding was derived based on only one single company, with the statistical method fitted to only eight quarterly observations (Fildes et al., 2009). By contrast, based on their study of products having intermittent demand, Syntetos et al. (2009) found that small adjustments to forecasts of zero demand are likely to be beneficial. Based on the above findings, the following two hypotheses were proposed:

H6b: The size of forecast adjustment is associated with the direction of forecast adjustment.

H6c: When adjustments are made, the size of forecast adjustments is positively associated with an improvement in accuracy.

3.7 Forecasting Performance Evaluation

The *ex ante* forecasting approach is clearly the most stringent and is also representative of the position of a practitioner producing forecasts; it was therefore adopted in this study. The definition of *ex ante* was obtained from Witt and Witt (1992). As illustrated by Figure 3.8, the model was first estimated using data for the period t_0 to t_1 and forecasts were made for the period t_1 to t_2. By using this approach, no advantage was taken of any information which would not have been available to a forecaster who was actually making the forecasts for t_1 to t_2 at point of time t.

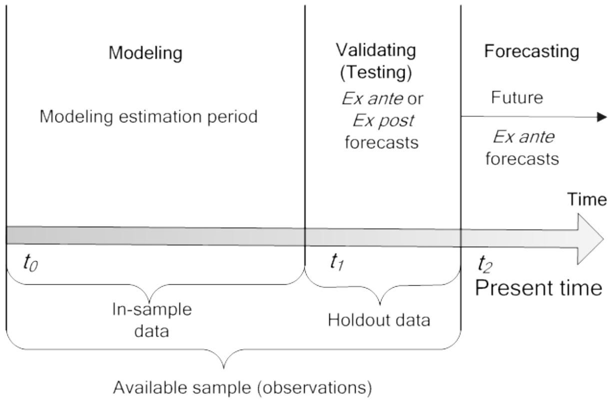


Figure 3.8 Time horizons of forecasts

Source: Song, Witt, and Li (2009, p. 183).

In order to decide which forecasting method is the best, it is necessary to have a yardstick by which to compare forecast accuracy. As suggested by Hatjoullis and Wood (1979, cited in Witt & Witt, 1992), "the principal difficulty in examining both questions 'how well do they forecast' and 'who forecasts best' is that there is no absolute yardstick against which forecasting performance can be judged" (p. 7). In this study, "accuracy" is used as the most important forecast performance judgment criterion. Forecast accuracy signifies the level of agreement between the actual values and the forecast values. Forecast accuracy is also regarded as the converse of forecast error, which is the difference between the actual value and the forecast. A small forecast error is an indication of high accuracy in forecasting. More rigorously, in this study, accuracy was constrained to be between 0 and 100% and defined accuracy as being equal to a maximum value of (100% − Forecast Error, 0).

In this section, important methodological issues related to testing the proposed hypotheses are discussed; these include the selection of error measures, the regression analysis of the forecasts, the measurement of data variability, the procedure for conducting a statistical analysis of the accuracy results, the tests for the bias and efficiency of judgmental forecasts, and the effects of the size of adjustments.

3.7.1 Error measures of forecast accuracy

The choice of an error measure can affect the ranking of forecasting methods (Armstrong, 2001b). The error measures selected should not only make sense to experts (face validity) but also produce findings that agree with other measures of accuracy (construct validity) (Armstrong, 2001b). The reason for employing more than one measure of error is that no single error measure has yet been shown to give an unambiguous indication of forecast accuracy (Armstrong, 2001e; Mathews & Diamantopoulos, 1986). The measures selected allow for the examination of the size as well as the directionality of forecast error in both absolute (volume) and relative (percentage) terms. Note that based on the assumption of a quadratic loss function, the squared error measure imposes higher penalties for large discrepancies between actuals and forecasts.

When the performance of forecasting methods needs to be compared across different time series, accuracy measures such as the mean squared error (MSE) and the mean absolute error (MAE) are inappropriate because there can often be major variations in the scale of the observations between series and so a few series with large values can dominate the comparisons (Goodwin & Lawton, 1999). Under such circumstances, unit free measures (e.g. MAPE) are more appropriate. Although the measurement of forecast accuracy is controversial, the use of absolute percentage error measures is now general practice within company settings (Fildes & Goodwin, 2007).

The following error measures were selected to evaluate the performance and accuracy of the forecasts in this study: R^2 value (correlation coefficient between predicted and observed values), the percentage better (PB) (than comparison forecasts), absolute percentage error (APE), MAPE, root mean squared percentage error (RMSPE), and Theil's U statistic (U statistic). Table 3.6 gives the formulae of the error measures employed in this study to assess forecasting performance, where n is the length of the forecasting horizon, A_t is the actual value at time period t, and F_t is the forecast made for period t.

Table 3.6 Measures of tourism demand forecast accuracy

Measure	Formula
Forecast error (E)	$E_t = A_t - F_t$
Percentage error (PE)	$PE_t = \left(\frac{A_t - F_t}{A_t}\right) \times 100$
Absolute percentage error (APE)	$APE = \frac{\lvert A_t - F_t \rvert}{A_t} \times 100$
Mean absolute percentage error (MAPE)	$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{\lvert A_t - F_t \rvert}{A_t} \times 100$
Root mean squared percentage error (RMSPE)	$RMSPE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\frac{A_t - F_t}{A_t})^2} \times 100$
Theil's U statistic (U)	$U = \sqrt{\frac{\sum_{t=1}^{n-1}\left(\frac{F_{t+1} - A_t}{A_t}\right)^2}{\sum_{t=1}^{n-1}\left(\frac{A_{t+1} - A_t}{A_t}\right)^2}}$

There are a number of reasons why it was decided to use MAPE as a measure of accuracy in this study. First, it is considered necessary to use a standardized measure to facilitate comparisons across the tourism demand flows being studied because they vary substantially in magnitude (Witt & Witt, 1992). Lewis (1982) stated that "[t]he MAPE is a most useful measure in comparing the accuracy of forecasts between different items or products since it measures *relative* performance" (p. 40). The MAPE is independent of the units used and can therefore be used to compare series with different units. To make the results more meaningful and comprehensible, the RMSPE was also applied where appropriate. Furthermore, MAPE was used because it was felt that it would be interesting to see if the conclusions regarding the ability of the various methods to forecast accurately differed when using different forecasting techniques. The use of MAPE allows for comparisons across different sizes of flows and does not penalize large errors, and thus it enabled a comparison of the accuracy of different tourism forecasts in this study. Furthermore, there is no justification for a nonlinear loss function (Larréché & Moinpour, 1983).

A smaller value of all of the measures (except for the U statistic) indicates that a better forecasting model produces predictions that are more accurate. The advantage of using the U statistic lies in the fact that it “allows a relative comparison of formal forecasting methods with Naive 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). When the value of the U statistic is less than one ($U < 1$), the forecasting technique being used is better than the Naive method, that is, the smaller the U statistic, the better the forecasting technique is relative to the Naive method. When the value of the U statistic is larger than one, it means that there is no point in using a formal forecasting method, since using a Naive method will produce better results. When the U statistic is equal to one, it indicates that the Naive method is as good as the forecasting technique being evaluated.

The value ranges for MAPE and U statistic are presented in Table 3.7. Lewis (1982) has suggested the following guidelines (see Table 3.7) for interpreting typical MAPE values; for example, if the MAPE of a model is less than 10 per cent, it is a highly accurate forecasting model.

Table 3.7 Interpretation of typical MAPE values and Theil's U statistics

MAPE (%)	Interpretation
< 10	Highly accurate forecasting
10-20	Good forecasting
20-50	Reasonable forecasting
> 50	Inaccurate forecasting
U Statistic	**Interpretation**
U = 1	Naive is as good as the forecasting model being evaluated.
U < 1	The forecasting model is better than Naive 1 approach, and this superiority increases as the U-statistic gets smaller.
U > 1	The Naive 1 model produces a more accurate forecast of the data series than the forecasting model under scrutiny, so there is no reason to employ it.

The U statistic is bounded between 0 and 1, with values closer to 0 indicating greater forecast accuracy. 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.

3.7.2 Regression analysis of the forecasts

An additional insight into the relative performance of forecasts can be obtained through the use of regression analysis, since this technique can show the degree of correspondence between the estimates (forecasts) and actual observations. Three pairs of regression analyses were performed in this study. In all cases, actual arrivals served as the dependent variable, the independent variable for each regression equation being, respectively, the initial statistical forecasts, the first round adjustments, and the second round adjustments.

Regression results are commonly evaluated based purely on R^2. As suggested by Armstrong (1985), an R^2 of one proves that the slope of the realized series is parallel with that of the forecasts. However, highly inaccurate forecasts can still achieve a high R^2; therefore, the slope and intercept should also be taken into account (Mathews & Diamantopoulos, 1989). A perfect forecast would yield an R^2 or unity, a slope of unity, and an intercept of zero. The adjusted coefficients of determination (R^2) obtained from regression analyses indicate a very close correspondence between actual arrivals and forecasts in all cases (Mathews & Diamantopoulos, 1986).

3.7.3 Tests for the bias and efficiency of judgmental forecasts

While accuracy is the most important property for a forecast, two further properties are also important: bias and efficiency. According to the studies by Ali, Klein, and Rosenfeld (1992), Harris (1999), and Lawrence, O'Connor, and Edmundson (2000), the bias and efficiency of judgmental forecasts can be investigated by fitting a regression model using the following equation:

$$PE_t = \alpha_0 + \beta_0 PE_{t-1} + \mu_t \tag{3.7}$$

where $PE_t = (A_t - F_t) / A_t$.

If a forecast is unbiased, the unconditional expectation of the forecast error should be zero (Harris, 1999). 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 or that forecasters are systematically overoptimistic as their forecasts are, on average, unfulfilled (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 or that forecasters are systematically overpessimistic as their forecasts are, on average, exceeded. 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).

The binomial test is useful for determining if the proportion of people in one of two categories is different from a specified amount. Although it is intended for categorical fields with only two categories, it can be applied to all fields by using rules for defining "success" (the expected proportion of records). The test is an exact test of the statistical significance of deviations from a binominal distribution of observations into two categories based on a specified probability parameter. By default, the hypothesized probability parameter for both groups is 0.5. The null hypothesis is that there is no difference between the two categories. By chance alone, half of the population are expected to be successes and half to be failures.

According to Harris (1999), an efficient forecast is defined as a forecast that can "optimally reflect currently available information, and is therefore associated with a forecast error that is unpredictable" (pp. 731-732). Nordhaus (1987) distinguished strong from weak efficiency and showed that it is very difficult to test strong efficiency in practice because "tests involve complete knowledge about the structure of the economy and access to private data that are not available to most econometricians" (p. 668). Harris (1999) also explained that to check a forecast for strong efficiency requires the forecast error to be uncorrelated with all of the information available at the time of the forecast. Testing weak efficiency, which requires that the forecast error is uncorrelated with the forecast itself, is thus recommended (Harris, 1999).

Under the null hypothesis that forecasts are weakly efficient, there will be no serial correlation in the errors from period to period (i.e. β_0 will be zero). This indicates that people will learn lessons from past errors when creating their new forecasts (Lawrence, O'Connor, & Edmundson, 2000). Rejecting the null that β_0 equals zero indicates that experts' forecasts do not fully incorporate the information contained in past forecast errors.

In addition to Equation (3.7) for testing weak efficiency, the following regression used by Harris' (1999) study is also applied in this study.

$$A_t = \alpha_1 + \beta_1 F_t + v_t \qquad (3.8)$$

The Wald test is used to test the joint null hypothesis that H_0: $\alpha_1 = 0$ and $\beta_1 = 1$. Accepting the null hypothesis suggests that the forecast is weakly efficient. If β_1 is significantly different from one, then conditional on the forecast itself, forecast error is predictable. If β_1 is significantly less than one, then the forecasters' estimates are too extreme, which means that high forecasts are related to high forecast errors while low forecasts are related to low forecast errors (Harris, 1999). If β_1 is significantly larger than unity, then forecasts are too compressed (Harris, 1999).

3.7.4 Measurement of data variability

In this study, it was of particular interest to examine the data variability that arises as a result of the special characteristics of tourism demand (e.g. seasonality, high sensitivity to external shocks, and policy impacts) as this is where judgmental adjustment is most needed. Data variability can be measured in a number of ways.

The simplest way is a judgmental assessment derived from an inspection of a scatter diagram of historical data. This method could possibly give forecasters a quick and intuitive means of assessing the predictability of unit arrival series. Adapted from Sanders and Ritzman's (1992) study, the coefficient of variation (CV) was used in this study to measure data variability in the arrival data series.

The CV is defined as the ratio of the sample standard deviation (*S*) to the sample mean ($\bar{x}$): CV = S / $\bar{x}$. It shows the extent of variability in relation to the mean of the population. For series with trend or seasonality or both, *S* can be replaced by the estimated standard error of an appropriate time series model. High values for this index indicate volatility and imprecise estimates about the trend line.

3.7.5 Statistical analysis of the accuracy results

Armstrong (1985) stated that "testing for statistical significance examines whether the superiority of a model is due to luck on the part of the researcher" (p. 356). It is necessary to conduct statistical tests to validate the comparisons made among different forecasts and to enhance the degree of generalizability of the findings. The objective of a statistical analysis is to examine the extent and nature of any differences in forecast error distributions between pairs of statistical forecasts and group forecasts.

To use parametric tests, the data should meet a number of assumptions: be normally distributed, have homogeneity of variance, and be continuous (the variables should be at least interval level of measurement or, if categorical, should have a minimum of seven categories) (Field, 2009; O'Neil, 2009). If the sample data seriously violate these assumptions, it will be safer to use nonparametric tests, which require fewer restrictions on the data sample.

Nonparametric tests are also known as assumption-free tests because they make fewer assumptions about the type of data on which they can be used (Field, 2009). The general assumptions of nonparametric tests include the following: independence of observations except when paired; few assumptions concerning the population' distribution; the scale of measurement of the dependent variable may be categorical or ordinal; the principle focus is either the ranking ordering or the frequencies of the data; sample size requirements are less stringent than for parametric tests (O'Neil, 2009). However, nonparametric tests should be used only when necessary as they sometimes reduce the ability to detect significant differences and thus have less power than their parametric counterparts (Field, 2009).

In this section, important issues regarding the statistical tests used for hypothesis testing in Chapter 5 are discussed. As shown in Figure 3.9, three steps were involved in selecting and conducting a statistical test to test a specific hypothesis. Details about each step are presented in the following sections.

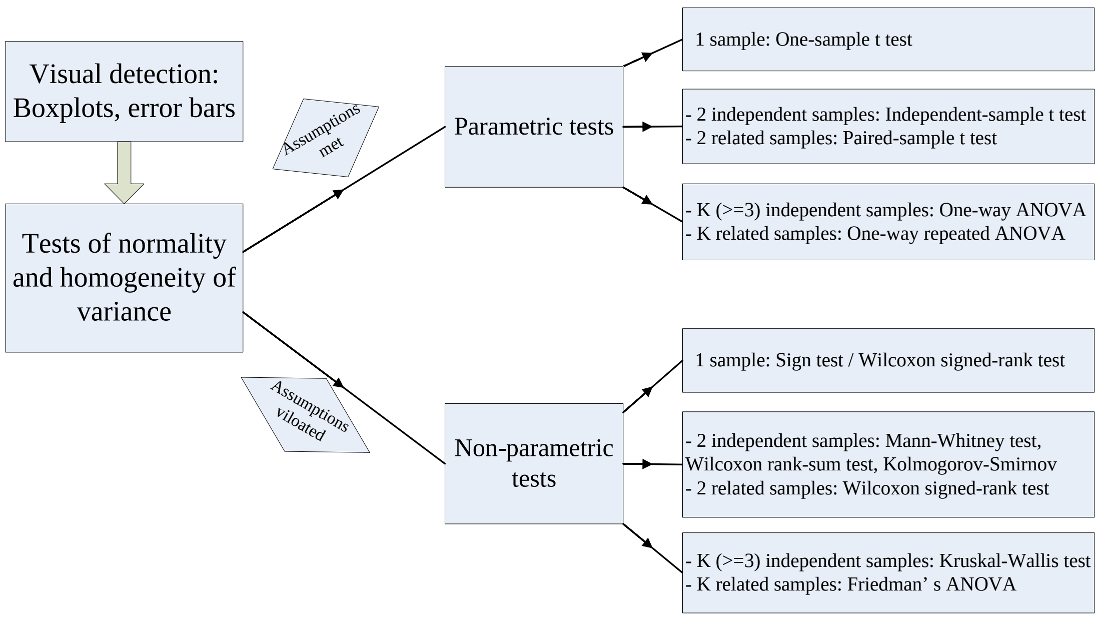


Figure 3.9 Flowchart of selecting statistical tests

(1) Visual detection

The first step of any statistical analysis is to graphically plot the data. One of the useful ways to check the data is to create a boxplot (also known as box-whisker diagram), which is an efficient method for displaying a five-number data summary − median, upper and lower quartiles, minimum and maximum data values (see Figure 3.10). In other words, a boxplot can provide information about the range, median, normality of the distribution, skewness of the distribution, and extreme cases within a sample.

The box plot shows a box encased by two outer lines known as whiskers. The box itself contains the middle 50% of a data sample − half of all cases are contained within it. With some exceptions (i.e. outliers or suspected outliers), the remaining 50% of the sample is contained within the areas between the box and the whiskers. Outliers are found in the form of points, circles, or asterisks outside of the boundaries of the whiskers; these are extreme values that deviate significantly from the rest of the sample and they can exist above or below the whiskers of the box plot.

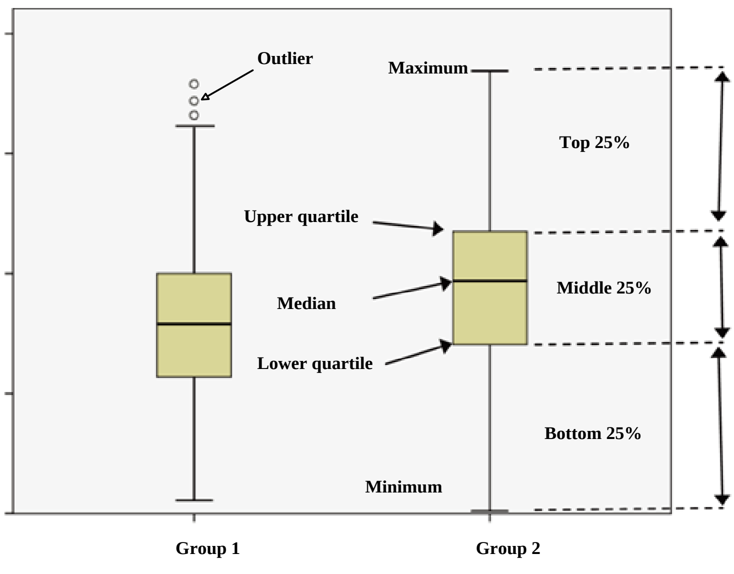


Figure 3.10 An example of a boxplot

Source: Adapted from Field (2009, p. 101).

The top edge of the tinted box in Figure 3.10 shows the value of the upper quartile (the 75th percentile of the data set), and the bottom edge indicates the value of the lower quartile (the 25th percentile of the data set). The line in the tinted box represents the value of the median. The location of the median line can also suggest skewness in the distribution: If the median line within the box is not equidistant from the edges, then the data is skewed. In addition, the location of the box within the whiskers can provide insight into the normality of the sample's distribution: If the box is shifted significantly to the low end, it is positively skewed; if the box is shifted significantly to the high end, it is negatively skewed.

In addition to the boxplot, an error bar graphically displays the 95% confidence interval of the mean for groups of cases. The boxes or circles in the middle of the error bar represent the mean score. The whiskers represent the 95% confidence interval.

(2) Normality and homogeneity tests

An assessment of the normality of data is a prerequisite for many statistical tests as normal data is an underlying assumption in parametric testing. Both the Kolmogorov-Smirnov (K-S) and Shapiro-Wilk (S-W) tests are used to see if the observed data fit a normal distribution. If a difference is detected, further tests can then be applied to establish the nature of the difference.

The S-W test is more appropriate for small sample sizes (< 50 samples) but can also handle sample sizes as large as 2,000, while the K-S test is more suitable for large samples. If the sample size is 50 or less, the S-W statistic should be used instead. If the *p* value (or *Sig.* in SPSS) of the test statistic is below 0.05, then the null hypothesis of no difference between the observed data distribution and a normal distribution is rejected. If the *p* value is greater than 0.05, then the data is normal. In other words, a *p* value less than 0.05 indicates that the data are nonnormal.

In addition to the normality tests, the homogeneity of variances should also be tested. One commonly used method to test such an assumption is the Levene's test, which tests the null hypothesis that the variances in different groups are equal. If the Levene's test is significant ($p < 0.05$), then equal variances are not assumed (heterogeneity); if the Levene's test is not significant ($p > 0.05$), then equal variances are assumed (i.e. homogeneity). As parametric tests are fairly robust to violations of homoscedasticity, the use of such tests is generally recommended unless the above tests for normality and homogeneity show strong departures.

(3) Parametric and nonparametric tests

The procedure shown in Figure 3.9 serves as a guide map for selecting an appropriate statistical test for the hypothesis testing in Chapter 5. Three examples are provided to illustrate how to conduct a statistical forecasting procedure according to Figure 3.9.

Case I: One sample

To test if the Round 2 group forecasts in this study were more accurate than the Round 1 group forecasts (*H3a*), the mean difference of MAPE and RMSPE were respectively calculated as *gapmape* and *gaprrmspe*. In total, 15 experts participated in both two rounds; therefore, 15 observations were included in *gapmape* and *gaprrmspe*. Table 3.8 presents the results from the K-S and S-W tests. The results show that the two series (i.e. *gapmape* and *gaprrmspe*) were both normally distributed as the *p* values were all above 0.05.

Table 3.8 Tests of normality for gapmape and gaprmspe

Measurement	K-S test			S-W test		
	Statistic	df	p	Statistic	df	p
gapmape	.181	15	.200	.907	15	.123
gaprmspe	.116	15	.200	.973	15	.901

Note: The variable *gapmape* is the difference between the MAPE value in R1 and R2, and the variable *gaprmspe* is the difference between the RMSPE value in R1 and R2.

Table 3.9 indicates that the assumption of homogeneity of variance was not violated as the p value of Levene's F statistic was greater than 0.05. Therefore, a parametric procedure (i.e. t test) that assumes normality and homogeneity can be applied. A one-sample t-test was conducted to ascertain whether accuracy had significantly improved over rounds. The null and alternative hypotheses were set as below:

H_o: *gapmape* (or *gaprmspe*) = 0, H_1: *gapmape* (or *gaprmspe*) > 0.

The above test aimed to examine the difference in accuracy in forecasting the different source markets rather than whether the average expert improved the statistical forecasts.

Table 3.9 Test of homogeneity of variance for gapmape and gaprmspe

Measurement		Levene Statistic	df1	df2	p
gapmape	Based on Mean	**1.158**	1	13	**.301**
	Based on Median	1.287	1	13	.277
	Based on Median and with adjusted df	1.287	1	10.826	.281
	Based on trimmed mean	1.118	1	13	.310
gaprmspe	Based on Mean	**1.157**	1	13	**.302**
	Based on Median	1.054	1	13	.323
	Based on Median and with adjusted df	1.054	1	10.866	.327
	Based on trimmed mean	1.135	1	13	.306

Case II: Two samples

The MAPE and RMSPE were used to test whether a difference in accuracy existed between the academic group and the industry group (*H5c*). Tests for normality and homogeneity of variance were first carried out to check the distribution for the two expert groups in each round in order to decide which type of statistical tests (i.e. parametric or nonparametric tests) should be applied.

Table 3.10 shows that in the first round data, the distributions for the industry and academic groups were found to be normal, whereas the second round data were nonnormal as three out of the four *p* values in the S-W test were less than 0.05 (two for MAPE and one for RMSPE). Note that due to the small sample size of the present study, the judgment on the normality test results was based on the S-W test. The Levene's test results in Table 3.11 show that the assumption of homogeneity was met for both MAPE and RMSPE over rounds at the 5% significance level. It was therefore decided to use the Mann-Whitney U-test to examine the accuracy difference between the two expert groups.

Table 3.10 Tests of normality for MAPE and RMSPE by expert group

Round		Group	K-S test			S-W test		
			Statistic	df	*p*	Statistic	df	*p*
R1	MAPE	Industry	.242	7	.200	.914	7	.426
		Academic	.174	11	.200	.955	11	.711
	RMSPE	Industry	.236	7	.200	.933	7	.579
		Academic	.162	11	.200	.985	11	.988
R2	MAPE	Industry	.279	6	.156	.760	6	*.025*
		Academic	.206	11	.200	.827	11	*.021*
	RMSPE	Industry	.319	6	.057	.809	6	.070
		Academic	.372	11	.000	.583	11	*.000*

Table 3.11 Test of homogeneity of variance for MAPE and RMSPE by expert group

Round			Levene statistic	df1	df2	*p*
R1	MAPE	Based on Mean	3.178	1	16	**.094**
		Based on Median	2.003	1	16	.176
		Based on Median and with adjusted df	2.003	1	14.286	.178
		Based on trimmed mean	3.177	1	16	.094
	RMSPE	Based on Mean	1.332	1	16	**.265**
		Based on Median	1.207	1	16	.288
		Based on Median and with adjusted df	1.207	1	14.890	.289
		Based on trimmed mean	1.340	1	16	.264
R2	MAPE	Based on Mean	.013	1	15	**.912**
		Based on Median	.000	1	15	.996
		Based on Median and with adjusted df	.000	1	14.668	.996
		Based on trimmed mean	.013	1	15	.912
	RMSPE	Based on Mean	.327	1	15	**.576**
		Based on Median	.158	1	15	.696
		Based on Median and with adjusted df	.158	1	11.616	.698
		Based on trimmed mean	.104	1	15	.751

Different from the Mann-Whitney U test (used for two independent samples), the Wilcoxon signed rank test is used for two related samples. These two tests are particularly appropriate for small sample sizes. The Wilcoxon signed-rank test (Siegel, 1956) is used to test for forecast accuracy differences between statistical forecasts and group forecasts in the case of nonnormality. The advantage of using this method is that not only can it determine the direction of any difference in forecast accuracy, but it can also take account of the magnitude of any difference between individual and combined forecasts (Song et al., 2009).

Case III: Three samples

The APE was used to test if there was a difference in accuracy among series with different data variability (*H5a*). The normality tests in Table 3.12 show that not all series followed normal distribution as there were two *p* values in the S-W test that were less than 0.05. Table 3.13 shows the results of the Levene's test. The assumption of homogeneity of variance was accepted as the significance of the Levene's tests was above 0.05 for all test series. Based on the above two test results, it was more appropriate to use the Kruskal-Wallis test.

If significance group difference is found, a post hoc procedure is then required to find out where the difference lies. As suggested by Field (2009), it is necessary to use a Bonferroni correction when applying nonparametric *post hoc* procedures. This means that instead of using 0.05 as the critical value for significance in each test, a critical value of 0.05 divided by the number of comparisons should be used to ensure that the Type I error does not build up to more than 0.05. In short, the main analysis using Mann-Whitney tests between pairs of conditions can be followed up in the *post hoc* procedures. However, a significant result can only be accepted when the significance of the test is below 0.05/number of tests. Therefore, in this study, all effects were reported at the 0.0167 (0.05/3) level of significance for the three comparisons.

As it is of interest to test whether the medians of the APE ascend or descend in the order specified by the data variability group, the Johckheere-Terpstra (J-T) test is used. The J-T test compares the medians of groups and checks whether there is an ordered pattern. The sign of the *z*-value indicates whether the trend of medians is ascending or descending. If it is positive, it indicates a trend of ascending medians; if it is negative, it indicates a trend of descending medians.

Table 3.12 Tests of normality for MAPE and RMSPE by data variability group

Measurement	CV group	K-S test			S-W test		
		Statistic	df	*p*	Statistic	df	*p*
APE_{SF}	Low	.234	5	.200	.912	5	.481
	Medium	.188	20	.062	.857	20	**.007**
	High	.261	5	.200	.922	5	.542
APE_{GF1}	Low	.375	5	**.020**	.750	5	**.030**
	Medium	.151	20	.200	.921	20	.102
	High	.334	5	.072	.878	5	.299
APE_{GF2}	Low	.237	5	.200	.903	5	.429
	Medium	.144	20	.200	.957	20	.478
	High	.318	5	.110	.890	5	.357

Table 3.13 Test of homogeneity of variance for MAPE and RMSPE by data variability group

Measurement		Levene Statistic	df1	df2	*p*
APE_{SF}	Based on Mean	1.870	2	27	.174
	Based on Median	1.470	2	27	.248
	Based on Median and with adjusted df	1.470	2	19.718	.254
	Based on trimmed mean	1.726	2	27	.197
APE_{GF1}	Based on Mean	2.019	2	27	.152
	Based on Median	1.586	2	27	.223
	Based on Median and with adjusted df	1.586	2	13.975	.240
	Based on trimmed mean	1.908	2	27	.168
APE_{GF2}	Based on Mean	2.420	2	27	.108
	Based on Median	1.875	2	27	.173
	Based on Median and with adjusted df	1.875	2	11.500	.197
	Based on trimmed mean	2.261	2	27	.124

(4) Effect sizes

To provide an objective measure of the importance of an effect, the effect size was used. An effect size is defined as “an objective measure and (usually) standardized measure of the magnitude of observed effect” (Field, 2009, p. 56). Effect sizes are useful because “they provide an objective measure of the importance of an effect” (Field, 2009, p. 57). A correlation coefficient of zero means there is no effect, and a value of one means that there is a perfect effect. Cohen (1992, cited in Field, 2009) made following suggestions to interpret the magnitude of an effect:

- r = 0.10 (small effect, or S): The effect accounts for 1% of the total variance.
- r = 0.30 (medium effect, or M): The effect accounts for 9% of the total variance.
- r = 0.50 (large effect, or L): The effect explains 25% of the total variance.

3.7.6 Effects of size of adjustments

The size of the adjustment was measured by its absolute size relative to the system forecast which is applied in Fildes et al.'s (2009) study. It is defined as:

$$SADJ_t = 100 \times \frac{|F_{2t} - F_{1t}|}{F_{1t}} \tag{3.9}$$

where F_{1t} is the forecast before adjustments, and F_{2t} is the forecast after adjustments. In addition to the raw error measures described earlier, a composite indicator of forecast improvement/degradation was constructed indicating how much closer to the actual arrival figure the forecast became as a result of experts' adjustments. The measure employed by Fildes et al. (2009) was applied in this study as follows:

$$\mathrm{IMP}_t = 100 \times \frac{AE_{1t} - AE_{2t}}{A_t} = 100 \times \frac{|A_t - F_{1t}| - |A_t - F_{2t}|}{A_t} \tag{3.10}$$

If IMP is positive this indicates that F_{2t} is more accurate than F_{1t} (i.e. the adjustment has improved forecasts). If IMP is negative, the opposite is true. The advantage of this measure is that it measures the improvement or degradation in MAPE introduced by the adjustment directly (Fildes et al., 2009).

3.8 In-depth Interviews Method

3.8.1 Overview of in-depth interview method

In-depth interviews are a qualitative research method for gathering information from individuals about their behaviour, opinions, feelings, and experiences (Longsfield, 2004). Such interviews are usually conducted face-to-face and involve one interviewer and one participant. Phone conversations and interviews with more than one participant also qualify as in-depth interviews. This technique is useful when researchers aim to obtain detailed information about a person's thoughts and behaviours or to explore new issues in depth (Boyce & Neale, 2006). The participants can be "members of the target audience or key informants, individuals who have special knowledge about the target audience, status among audience members, access to important information, or a willingness to share their knowledge and skills" (Longsfield, 2004, p. 2).

Although used less often than focus group discussions, in-depth interviews can be used if the potential participants may not be included in, or feel uncomfortable talking openly in a group or when researchers aim to distinguish individual (as opposed to group) opinions about a subject. As summarized by Longsfield (2004), in-depth interviews can be used for a variety of purposes (see Table 3.14); for example, during performance monitoring, in-depth interviews can provide programmes with participants' feedback about intervention efforts and identify areas for further improvements. In-depth interviews can also be used during evaluation to clarify survey findings and solicit additional feedback from target audience members, key informants, or project implementers.

Table 3.14 An overview of in-depth interviews

Purpose of in-depth interviews:	
Explore a relatively unknown behaviour.	Examine a sensitive study topic.
Inform campaign/program development (pre-testing).	Obtain information from knowledgeable informants.
Study complex behaviours and motivations.	Uncover local terms related to a topic.
Work with geographically dispersed informants.	Learn the "how" and "why" behind behaviour.
Reveal images, language, concepts, and packaging that appeal to audiences (concept testing).	Obtain information that might be influenced by peer pressure during focus group discussions.
Generate new ideas for a program.	Develop language and survey content.
Generate hypotheses for future research.	Improve project implementation.
Clarify survey findings.	
Advantages and Disadvantages of in-depth interviews:	
Advantages	*Disadvantages*
Uncover valuable insights, and find out "the real story" from the people in the know.	Quality of data depends on the interviewer and quality of transcription skills.
Respondents are most likely to open up on a one-on-one basis, and are less influenced by peers than the focus group method.	*Interviewing requires a high level of training and skill*: It is important to have well-trained, highly-skilled interviewers conducting this type of interview. Using less skilled interviewers increases the possibility of bias.
Skilled interviewers are able respond to questions and probe for greater details. Questions can be added or altered in real-time if needed.	A small sample size.
Require less time for data collection.	Require a great deal of time for data analysis.
Cost efficient.	Less cost-efficient than focus groups.
Appropriate when access to groups is limited.	More interviews needed than a focus group method to reach as many participants.
Provide details about sensitive information, including personal experiences, views, and behaviour.	Inappropriate for determining programme "impact," social norms and trends.
Provide confidential atmosphere for informants.	It cannot be generalized to larger audiences.
Mobile method	Researcher has little control over the environment since interviews may take place in a variety of settings.

Source: Adapted from Boyce and Neale (2006, p. 4), Longsfield (2004, p. 4), and The Wallace Foundation (2012, p. 4).

Table 3.14 summarizes the advantages and disadvantages of using in-depth interviews. The primary advantage of in-depth interviews is that participants can provide detailed information about their personal experiences, views, and behaviour.

In addition, in-depth interviews provide a confidential atmosphere in which participants can share sensitive information. Such a private setting also means that peers do not influence participants' responses to the study topics. Compared to focus groups, in-depth interviews are more time intensive due to the time it takes to conduct interviews, transcribe them, and analyse the results. Furthermore, if they choose to use in-depth interviews, researchers have to conduct more sessions to obtain as many different perspectives as possible.

Despite being time consuming, in-depth interviews can allow researchers to collect a great deal of information in a short period (e.g. a few weeks). In addition, interviews can permit access to audience targets when groups are difficult to coordinate or contact and can also allow researchers to work with geographically dispersed informants. However, there are a few limitations and pitfalls of in-depth interviews, and these are described in Table 3.14.

Given the length of each interview and the associated costs, the number of in-depth interviews is usually small. The non-probability sampling method is usually applied to recruit participants, which suggests that such a sample is not representative of a larger population. As a result, the results from in-depth interviews are only suggestive of trends among the informants and usually generalizations cannot be made from the results because small samples are chosen and random sampling methods are not used.

The number and composition of in-depth interviews depend on the study objectives, the characteristics of the target population, and the study locations (Longsfield, 2004). There is no standard number of interviews, but it is much more common for as few as 10 to 15 interviews to be conducted (Boyce & Neale, 2006). Longsfield (2004) found that for formative research, it is appropriate to use 6 to 20 interviews; however, fewer interviews are needed when they are supplemented with other data or when research objectives are very limited. Longsfield (2004) further suggested that researchers should, as a minimum, conduct two interviews for each type of informant in the target audience; the rationale behind this is that conducting at least two interviews for each type of informant will "ensure that if one interview does not go well, the research team still has another interview with the same type of informant from which to collect data" and that it "permits researchers to confirm the reliability of study data" (Longsfield, 2004, p. 14). It should be noted that the general rule on sample size for interviews is that "when the same stories, themes, issues, and topics are emerging from the interviews, then a sufficient sample size has been reached" (Boyce & Neale, 2006, p. 4).

3.8.2 Justification for using in-depth interviews

Error measures and statistical tests provide quantitative information to evaluate the forecasting performance of statistical and judgmental forecasts. However, in this study, additional information, such as how the Delphi participants integrated their knowledge and expertise into their judgmental forecasts, what underlying assumptions lay behind their adjustment process, and how they used the forecasting support system (HKTDFS) to assist with their forecasting, could not be obtained through the above quantitative analysis. To obtain valuable insights into the experts' adjustment behaviour, explore the reasons for the accuracy improvements from the proposed integrative framework, and investigate the causes of the biases and inefficiency in the judgmental adjustments, in-depth interviews were conducted among those experts who participated in the main Delphi surveys.

In this study, the purposes of conducting in-depth interviews were to (a) explore the underlying assumptions that individual experts made during their Delphi forecasting; (b) investigate the possible reasons why the proposed integrative forecasting approach could produce accurate forecasts; (c) discover the experts' views on how the forecasting system could aid their judgmental adjustments (e.g. the presentation of graphs and tables; the provision of historical series, etc.); (d) provide suggestions to further improve the forecasting performance of the current integrative forecasting framework in tourism; and (e) make recommendations to further enhance the forecasting ability of the HKTDFS.

There were several reasons why in-depth interviews had advantages over the focus group method in terms of achieving the above objectives. First, the participants were from competing organizations and may not have been comfortable talking openly in a group. Second, the participants may have used forecasts in different ways and for different purposes and may have had different preferences, and so one person's experience and needs would not be of interest to the others. Third, the informants were more likely to share their true thoughts and experiences in a confidential individual setting. Fourth, the participants' views or opinions may have been distorted because they wanted to impress the others or to "go along with the crowd". Last but not least, it would have been logistically difficult to get the participants in one room at one time as both the industry practitioners and the academic researchers had very tight schedules and some of them were not in Hong Kong during the interview period (June-July 2012).

3.8.3 Procedure for conducting in-depth interviews

The five phases involved in conducting in-depth interviews are presented in Figure 3.11: (a) planning, (b) writing an interview guide, (c) conducting interviews,

(d) analysing the data, and (e) reporting. More details regarding each step are provided in the remainder of this section.

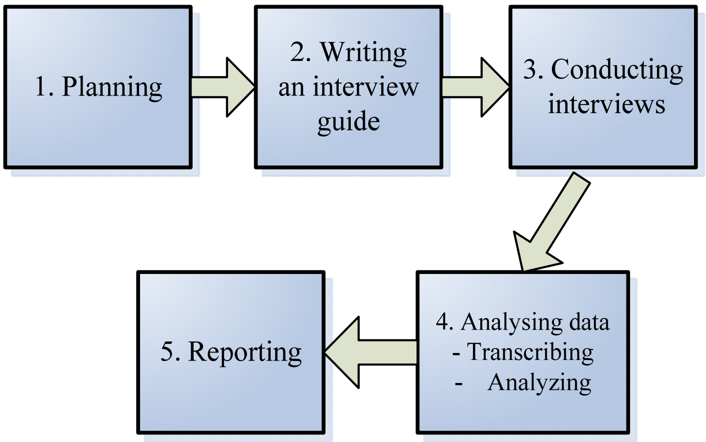


Figure 3.11 Steps involved in conducting the in-depth interviews

(1) Planning

The first phase in conducting in-depth interviews involves developing a recruiting strategy to identify stakeholders who could be involved in this study and to figure out how to find these people.

Who is the target audience?

It is often difficult to find informants who meet the eligibility criteria and feel comfortable discussing issues with researchers. Friends and family members of interviewers are not eligible candidates since their familiarity with the topic and the researcher may bias study results (Longsfield, 2004). According to Spradley (1979), good informants are those people who know the local culture, are involved with the study topic, can share first-hand experience of the study topic, have adequate time to devote to an interview, are impartial, and have not already analysed the study topic from an outsider's perspective.

For this study, the potential candidates for interview were selected from those who had participated in the main Delphi survey of this project. The issues and concerns raised in the early round of interviews (i.e. in the pretesting stage) informed the interview guides used in the subsequent round of interviewing.

Selection of participants (or sampling)

Judgment sampling was used to select the participants for the in-depth interviews. A research sample was created based on those panellists who had participated in either the first or second round of the Delphi surveys. The reason for the inclusion of these Delphi panellists was that they were the most informed and had the most to contribute to the study topic. The potential respondents were contacted in two stages. At the initial point of contact, invitation letters were sent out to a total of 21 Delphi experts to check their willingness to participate in the interviews. In the subsequent stage, only those who agreed to be interviewed were tracked. Emails were then sent to them to explain the purposes of the interview and to schedule a time and place to conduct the interview.

(2) Writing an interview guide

Before developing an interview guide, it was necessary to determine what type of interviews would be suitable for this project. Structured interviews are most likely to be appropriate when conducting interviews by telephone, face-to-face interviews in informants’ households, intercept interviews, and interviews associated with survey research (Denzin & Lincoln, 2000). Structured interviews were used in this study because they use predetermined questionnaires, which can help to probe knowledge, attitudes, beliefs, and practices, and can allow for comparisons among different informants.

An in-depth interview guide is a method for guiding the administration and implementation of interviews to help the interviewer focus on topics that are important to explore and to ensure consistency across interviews with different respondents and thus increase the reliability of the findings. A comprehensive interview guide is essential for conducting good interviews. Guides should include appropriate sections, contain key questions that answer study objectives, and meet the needs of data users (Longsfield, 2004). The contents of an interview guide are mostly determined by the research objectives (The Wallace Foundation, 2012). A common mistake in developing guides is to ask too many questions, resulting in an unwieldy guide, long interviews, and compromised data. When guides are too long, interviewers will not have sufficient time to fully explore the designed topics and will not get the full benefit of using an in-depth interview. Boyce and Neale (2006) suggested that "there should be no more than 15 main questions to guide the interview, and probes should be included where helpful" (p. 5). In addition, when developing guides, researchers should always reflect on their analysis plan and the appropriate report format (Longsfield, 2004).

For this project, the interview guide (see Appendix B) was developed and formatted into four sections based on the two studies of Longsfield (2004) and Guion, Diehl, and McDonald (2011): (a) *factsheet*, which was used to record the time, date, and place of the interview; special conditions that may affect the interview; and demographic information about the respondent being interviewed; (b) *introduction*, which explained the purposes of the interview, made informants feel comfortable, and set the tone for the rest of the discussion; (c) *interview questions*, or the heart of the guide, which included the key questions directly related to the study objectives and follow-up or probing questions to explore specific aspects of an issue; and (d) *wrap-up*, which summarized the discussion, thanked the informant for their participation, and asked if there was anything else the informant would like to add. To reflect the different nature of forecasters, two sets of interview questions were developed for the two groups of target respondents, namely industry and academic groups (see Appendix B).

After the interview guide was drafted, it was sent to two academic experts (with expertise in econometric modelling and judgmental forecasting respectively) for review and feedback. The guide was revised based on the two experts' comments and suggestions. Subsequently, pretesting was conducted using two Delphi experts to check the appropriateness of the question format, the length of the interview, the clarity of the contents, and the language. The pretesting questions are listed in Appendix C.

(3) Conducting the interviews

Solid preparations should be made before formal interviews. During planning, the participants' access to transportation, time schedule, and personal preferences for interview site should be taken into consideration. In the current study, a checklist was made to minimize the possibility of forgetting necessary items (see Appendix D) and appointments to conduct the interviews were made in advance.

To conduct a successful in-depth interview, it is important to begin with a brief introduction to the study, the interviewer, and the informant. This component offers an opportunity to establish a good rapport with respondents by making them feel welcome, thanking them in advance for their participation, and laying the foundation for a successful interview (Longsfield, 2004). If audio recording is required, it is important to first obtain the respondent's permission and to test the equipment to make sure that it is working properly. The main responsibility of the interviewer is to listen and observe until all of the important issues on the interview guide have been explored. Appendix D shows the guidance used in this study on how to promote discussions during interviews.

(4) Analysing the data

Transcribing. Transcribing involves "creating a verbatim text of each interview by writing out each question and response using the audio recording" (Guion, Diehl, & McDonald, 2011, p. 3). The transcription should also include the interviewers' side notes about their impressions of the interview, the respondent's nonverbal behaviour, and the rapport between them and the interviewees.

Analysing. The analysis of in-depth interviews is called *contextual analysis*, which requires the thorough reading and coding of interview transcripts (Longsfield, 2004). It requires researchers to identify common trends and patterns/themes that appeared among the informants' responses. The analysis process also helps to search for explanations of behaviour and supporting quotes that are integrated into the final report (Longsfield, 2004). One common strategy, namely organizing transcripts by question, was applied in this study. Specifically, with this strategy, questions were used to organize the data analysis, in essence synthesizing the answers to the questions that were proposed in the interview guide.

(5) Reporting

Finally, it is important to disseminate interview findings to interviewees through a written report and to solicit feedback if possible. In general, the report should provide a general description of the study sample, highlight patterns and recurring themes across interviews, address informants' concerns, cover conflicting information, present unexpected findings, and propose suggestions for shaping future work in question. Apart from the evaluation results, background information and the methodology of the study, together with any supporting materials (e.g. the interview guide), should also be included in the report (Boyce & Neale, 2006). When respondents see the information being used, they are more likely to participate in future data-collection efforts.

3.8.4 Profile of the respondents

In this study, actual contact was made with 21 experts through both emails and phone calls, and 14 agreed to participate in the in-depth interviews. The panel was composed of five industry experts (three from tourist attractions and two from government) and nine academic experts. Table 3.15 gives the accuracy ranking of the panellists according to their Delphi forecasting performance. The two informants with star symbols were the experts who did not directly participate in the Delphi survey but did participate as team members. They were also regarded as interviewees who were qualified to share their insights and viewpoints on the study topic.

Table 3.15 Composition of in-depth interview participants

Accuracy ranking	Category	Pretesting
1	Tourist Attractions	No
*	Tourist Attractions	No
2	Academic	***Yes***
3	Government	No
*	Government	No
4	Academic	No
6	Academic	***Yes***
8	Academic	No
10	Academic	No
11	Academic	No
14	Tourist Attractions	No
16	Academic	No
17	Academic	No
18	Academic	No

3.8.5 Timeline

The timeline for completing the in-depth interview is presented in Figure 3.12. It took one week (from June 1 to June 7) to outline the study objectives and another three weeks to draft and finalize the interview guide. Recruitment started 2 weeks before the main interviews were conducted. Invitation letters were sent out via email to the panellists involved in the main Delphi surveys, which were conducted between June and July 2011. Pretesting was conducted after the interview guide had been drafted and tested by two academic experts. Subsequently, the interview guide was refined according to the feedback and comments obtained from two interviews in the pretesting phase. A reminder email was sent out 1 day before the scheduled interview. Only one interview was scheduled per day, and it took another 1 or 2 days to translate and transcribe the interview. Data coding and analysis took about six weeks, and report writing took 1 week to complete.

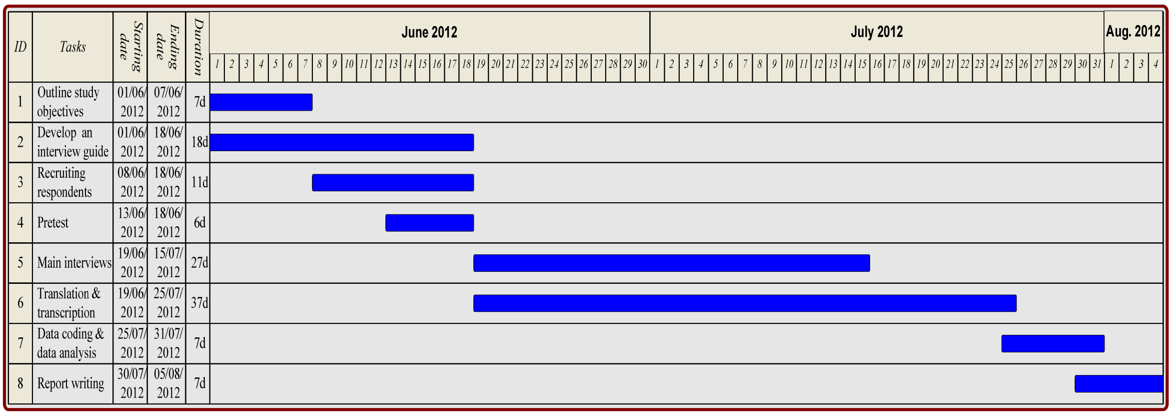


Figure 3.12 Gantt graph for conducting in-depth interviews

3.9 Chapter Summary

This chapter has described the necessary methodology issues to be considered when evaluating forecasting performance and exploring experts' judgmental forecasting behaviour in Chapter 5. The contents of this chapter were organized according to the progress of the project: selecting forecasting variables, preparing data, developing statistical forecasting models to generate quantitative forecasts, making judgmental adjustments, evaluating forecasting performance, and conducting in-depth interviews.

Rationales and justifications regarding the choice of the selected model or method in this study were provided for each step of the analysis. The decisions on determining the statistical forecasting models and selecting the dependent variable (i.e. visitor arrivals) and its key determinants (i.e. income, own price, substitute price, dummy variables) were made based on a review of the existing tourism demand forecasting literature. Justifications for using the HKTDFS as the Web-based forecasting support system to structure the judgmental forecasting procedure and for choosing the Delphi forecasting method to aggregate experts' judgmental adjustments were made to enhance the validity and reliability of the study. A group of error measures and statistical tests were employed to test the proposed research hypotheses, and these can help to evaluate the effectiveness of the statistical and judgmental forecasts. Despite the emphasis on the statistical testing of the forecasting results in this study, the forecast evaluation methods tended to be based on indications from the existing literature and intuitive ideas and the associated statistical techniques were mostly straightforward.

Apart from the evaluation of accuracy, regression analysis was used to investigate whether the statistical and judgmental forecasts were unbiased and efficient. Lastly, in-depth interviews were conducted to follow up the main analysis in order to give the experts' views and perceptions of the forecasting procedure.

Chapter 4 : A Pilot Study[8]

4.1 Introduction

The methods introduced in the preceding chapter were served as the guide to conduct this. As clarified in the preceding chapter, this study used a group-based forecasting support system− Delphi survey conducted via the HKTDFS − to systematically integrate experts' group judgments into the statistical forecasts. The forecasting adjustment procedure was employed in the HKTDFS Statistical Adjustment module. Before formally launching the main Delphi survey, it is necessary to test the reliability of this module. An experiment was thus carried out with the involvement of postgraduate students and staff from the School of Hotel and Tourism Management at The Hong Kong Polytechnic University. This chapter presents results from this experimental survey. In Section 4.2, a summary of the pilot survey is provided. In Section 4.3, forecast accuracy is evaluated by three error measures. Section 4.4 presents results from a feedback survey after the experiment to further improve the functional ability of the system. Section 4.5 concludes the chapter by summarizing the findings.

4.2 Descriptive Summary

This two-round survey was undertaken over the period June 5–11, 2011. All participants were asked to self-rate their level of expertise in tourism demand forecasting on a 7-point Likert scale, ranging from 1 ("*very little*") to 7 ("*excellent*"); the 16 participants who were included in the survey had a mean self-rating score of

[8] Parts of this chapter was published in Song, Gao, and Lin (2013).

4.07. Of these 16 participants, 19 per cent rated themselves as having very little expertise, whereas 6 per cent rated themselves as having level 6 experience (see Figure 4.1). About 38 per cent rated themselves as level 5, 32 per cent fell in levels 3 and 4, and the remaining 30 per cent in other levels.

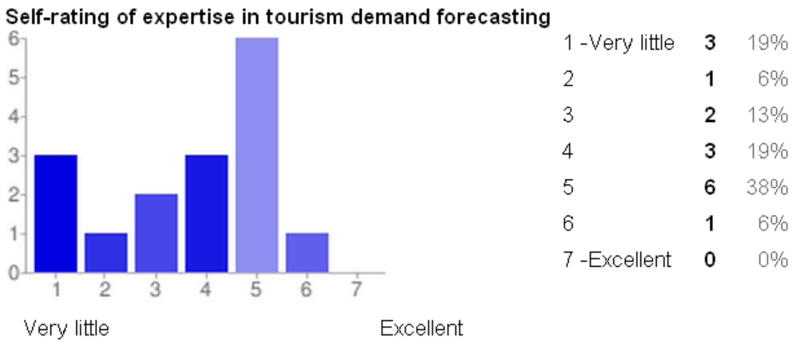
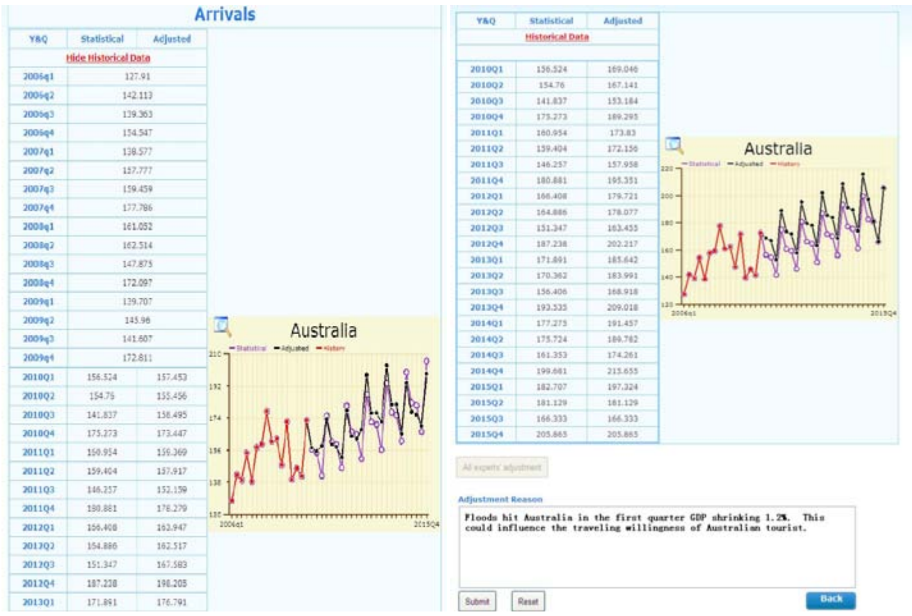


Figure 4.1 Self-rating of expertise by participant

Before carrying out the forecasting tasks, the participants were asked to read the introductory document, to familiarize themselves with the procedure for using HKTDFS. Students who confirmed their desire to participate were each assigned a user ID and a password to enable them to access the HKTDFS. Participants were invited to make adjustments to the quarterly forecasts of visitor arrivals from three short-haul markets (China, Taiwan, and Japan) and three long-haul markets (the USA, the UK, and Australia) served by the Hong Kong tourism industry over the period 2010Q1–2015Q4. Statistical forecasts were produced by the ARDL-ECM method using the sample 1985Q1–2009Q4. Participants were asked to consider the effects of two special events which were not taken into account by the ARDL-ECM, namely the 2011 earthquake in Japan and the launch of the Beijing-Shanghai high speed railway. Annual projections of the real GDP growth rates and the exchange rates obtained from IMF for the six source markets were also provided to the participants.

One major characteristic of the HKTDFS is that it is user-oriented, in that it allows the user to make a wide variety of interventions through judgmental adjustments, incorporating the effects of special events. Two options were available to users: making annual or quarterly adjustments. In addition, users could also make adjustments for different forecasting periods (see Figure 4.2). The statistical forecasts for the period 2010Q1–2015Q4 were presented, together with the individual forecasts in both tabular and graphical form (see Figure 4.3 (a)). Individual forecasts were set to match the statistical forecasts. Users could choose different output options (with both historical data and forecasts or with forecasts only), as shown in Figure 4.3 (a). Moreover, users' justifications for their adjustments could be stored within the HKTDFS for subsequent reference (see Figure 4.3 (b)).

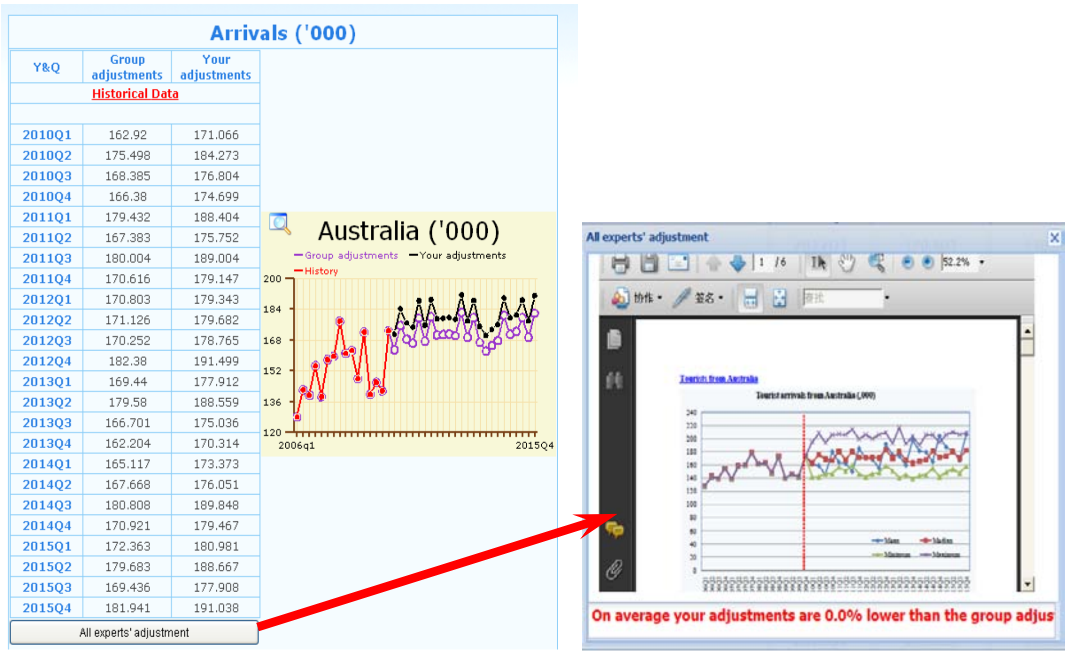
Figure 4.2 Options for adjustments



(a) (b)

Figure 4.3 Screen shots from the HKTDFS (R1)

In the first round, positive responses were received from 56.5 per cent of the selected panellists. In the second round, the median forecasts from Round 1 were presented as the baseline forecasts (see Group Adjustments in Figure 4.4 (a)). The participants were then required to verify and adjust these group forecasts. The system also allows access to the statistical summary report and written justifications (see Figure 4.4 (b)), with a view to informing the participants of the adjustments made by other participants and the reasons for these adjustments.



Y&Q	Group adjustments	Your adjustments
2010Q1	162.92	171.066
2010Q2	175.498	184.273
2010Q3	168.385	176.804
2010Q4	166.38	174.699
2011Q1	179.432	188.404
2011Q2	167.383	175.752
2011Q3	180.004	189.004
2011Q4	170.616	179.147
2012Q1	170.803	179.343
2012Q2	171.126	179.682
2012Q3	170.252	178.765
2012Q4	182.38	191.499
2013Q1	169.44	177.912
2013Q2	179.58	188.559
2013Q3	166.701	175.036
2013Q4	162.204	170.314
2014Q1	165.117	173.373
2014Q2	167.668	176.051
2014Q3	180.808	189.848
2014Q4	170.921	179.467
2015Q1	172.363	180.981
2015Q2	179.683	188.667
2015Q3	169.436	177.908
2015Q4	181.941	191.038

(a) (b)

Figure 4.4 Screen shots from the HKTDFS (R2)

4.3 Evaluation of Forecast Accuracy

Three error measures, the APE, MAPE, and RMSPE, were used to evaluate the forecast accuracy. Out-of-sample forecast errors were generated for the period 2010Q1–2011Q2. As anticipated, the judgmentally adjusted forecasts were more accurate than the statistical forecasts (i.e. the average MAPEs decreased from 8.86% to 8.02% and the average RMSPEs from 10.41% to 9.33%). Furthermore, the accuracy also increased over the rounds in terms of both the MAPE and RMSPE (see Table 4.1). Even though the forecast adjustments improved the forecast accuracy on average, the level of improvement varied across source markets: the accuracy improved for all three short-haul markets and one long-haul market (the USA), but decreased in Australia and Taiwan (see Table 4.1). In other words, the overall improvement came largely from the contribution of the three short-haul markets.

Table 4.1 MAPE and RMSPE

Country/Region	**MAPE (%)**			**RMSPE (%)**		
	SF	GF1	GF2	SF	GF1	GF2
China	16.19	11.91	11.29	17.64	13.78	13.03
Taiwan	10.74	8.80	9.02	12.20	10.02	10.26
Japan	8.45	7.91	7.87	11.61	10.49	10.65
Australia	2.94	3.30	4.88	4.64	4.26	5.29
UK	7.58	11.54	10.47	8.95	12.47	11.55
USA	7.25	4.69	4.39	7.41	4.93	4.64
Mean $_{Short\text{-}haul}$	11.79	9.54	9.39	13.82	11.43	11.31
Mean $_{Long\text{-}haul}$	5.93	6.51	6.58	7.00	7.22	7.16
***Mean** $_{Total}$*	***8.86***	***8.02***	***7.99***	***10.41***	***9.33***	***9.24***
Error reduction (%)	GF1-SF	GF2-SF	GF2-GF1	GF1-SF	GF2-SF	GF1-GF2
China	–4.28	–4.90	–0.62	–3.86	–4.61	–0.75
Taiwan	–1.94	–1.72	0.22	–2.19	–1.94	0.24
Japan	–0.54	–0.58	–0.04	–1.12	–0.96	0.16
Australia	0.35	1.94	1.59	–0.38	0.66	1.03
UK	3.95	2.88	–1.07	3.52	2.59	–0.92
USA	–2.57	–2.87	–0.30	–2.48	–2.77	–0.29
Mean $_{Short\text{-}haul}$	*–2.26*	*–2.40*	*–0.14*	*–2.39*	*–2.50*	*–0.12*
Mean $_{Long\text{-}haul}$	*0.58*	*0.65*	*0.07*	*0.22*	*0.16*	*–0.06*
Mean	***–0.84***	***–0.87***	***–0.04***	***–1.08***	***–1.17***	***–0.09***
Percentage reduction (%)	GF1-SF	GF2-SF	GF2-GF1	GF1-SF	GF2-SF	GF2-GF1
China	–26.44	–30.25	–5.18	–21.87	–26.13	–5.45
Taiwan	–18.09	–16.03	2.52	–17.92	–15.93	2.43
Japan	–6.45	–6.89	–0.48	–9.62	–8.27	1.48
Australia	12.05	65.95	48.11	–8.15	14.16	24.28
UK	52.09	37.98	–9.28	39.27	28.97	–7.39
USA	–35.39	–39.52	–6.39	–33.46	–37.41	–5.94
Mean $_{Short\text{-}haul}$	–16.99	–17.72	–1.05	–16.47	–16.78	–0.51
Mean $_{Long\text{-}haul}$	9.58	21.47	10.81	–0.78	1.91	3.65
Mean	***–3.71***	***1.87***	***4.88***	***–8.62***	***–7.44***	***1.57***

Note: SF, GF1, and GF2 represent the econometric (statistical) forecasts, group 1 and 2 forecasts, respectively.

The MAPEs of the statistical forecasts for the long-haul markets were far smaller than those of the short-haul markets; however, the forecasting performance deteriorated for Australia and the UK after judgmental adjustments. When the statistical forecasts were highly accurate, judgmental adjustment seemed to have little impact on the accuracy, or even reduced it. One possible explanation for this finding may lie in the capacity of econometric models to make accurate extrapolations and identify established patterns and existing relationships, and thus produce highly accurate forecasts (e.g. all MAPEs for the three long-haul markets were less than 8 per cent). Under such conditions, judgmental revisions of the statistical forecasts may tend to overreact to fluctuations in the arrival series (Sanders & Ritzman, 1995). Another factor which contributed to the forecasting performance is the volatility of the time series being forecast.

As shown in Figure 4.5, the historical data series of visitor arrivals for the short-haul markets are more volatile (or less stable) than those in the long-haul markets. This is reflected in the coefficient variation for the six markets: 1.1 (China), 0.39 (Taiwan), 0.32 (Japan), 0.33 (the UK), 0.25 (the USA), and 0.40 (Australia). Sanders and Ritzman (2004) suggested that less emphasis should be placed on contextual knowledge when making combination forecasts if the variability is low.

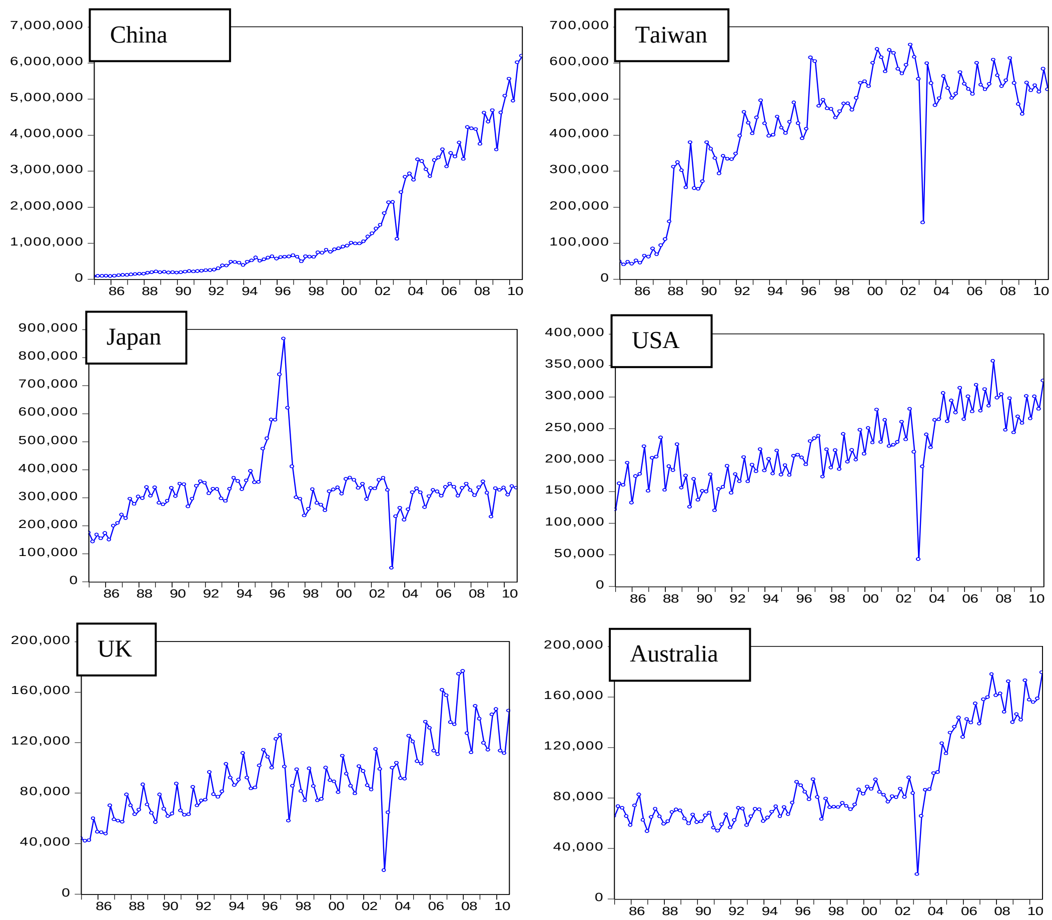


Figure 4.5 Historical trends of visitor arrivals

Error reductions between each of the pairs of forecasts, namely the statistical forecasts and the two rounds of judgmental forecasts, were computed and are shown in Table 4.1. A negative change in either the MAPE or RMSPE indicates an improvement in accuracy. The greatest improvement in accuracy over the statistical forecasts was in the predictions of visitor arrivals from the USA, followed by those from Mainland China. The big improvement in forecasting US visitors to Hong Kong may be due to the provision of useful feedback from participants. For example, one panellist pointed out that “with some signs of recovery from the global financial crisis, arrivals from the USA can improve faster than the statistical trend”, and this turned out to be the case after comparing the statistical forecasts with the actual arrivals over the period 2010Q1–2011Q2. As was discussed earlier, the series of Chinese visitor arrivals was the most unstable one, with the largest coefficient of variation (1.1). Judgmental inputs for these series could significantly improve the forecast accuracy (Sanders & Ritzman, 1995). The smallest improvement over the statistical forecasts made by judgmental interventions was in predicting visitor arrivals from the UK However, as shown in Table 4.1, the greatest improvement in accuracy from round 1 to round 2 was achieved in the case of the UK (9.28% and 7.39% reductions in MAPE and RMSPE respectively), followed by the USA and China.

A more detailed analysis of the performance statistics (see Table 4.2) reveals that the group forecasts from the second round were more accurate than either the statistical forecasts or the group forecasts from the first round. Taking China as an example, the APEs of the three sets of forecasts were calculated for each quarter between 2010Q1 and 2011Q2. The cumulative frequencies of negative error differences between the two forecasts, measured by the APE, are given as percentages. As Table 4.2 shows, an improvement in forecast accuracy as a result of using the combination method (versus statistical forecasting alone) was observed in all quarters in the cases of China, Taiwan and the USA; however, the accuracy of the combination method decreased in the case of the UK.

Table 4.2 Forecast performance evaluated by APE

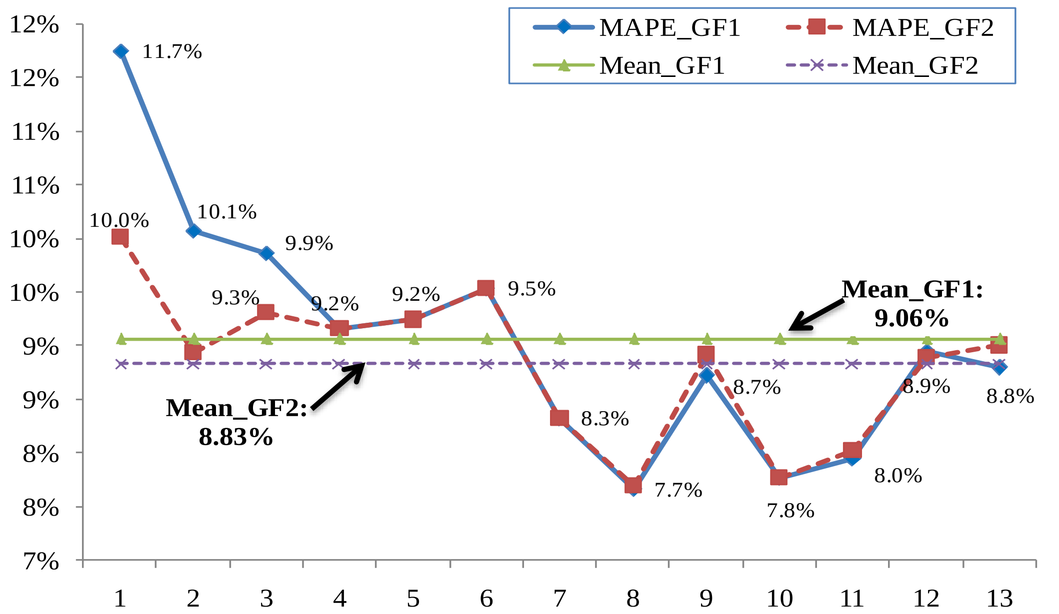
Country /Region	*Quarter*	APE_{SF} (a)	APE_{GF1} (b)	APE_{GF2} (c)	$\%_{(b-a<0)}$	$\%_{(c-a<0)}$	$\%_{(c-b<0)}$
China	2010Q1	11.38	7.58	7.60	100	100	83
	2010Q2	21.96	17.73	17.32			
	2010Q3	15.22	11.34	10.87			
	2010Q4	6.94	2.29	1.85			
	2011Q1	13.44	8.98	8.48			
	2011Q2	28.18	23.51	21.62			

Table 4.2 Forecast performance evaluated by APE (Continued)

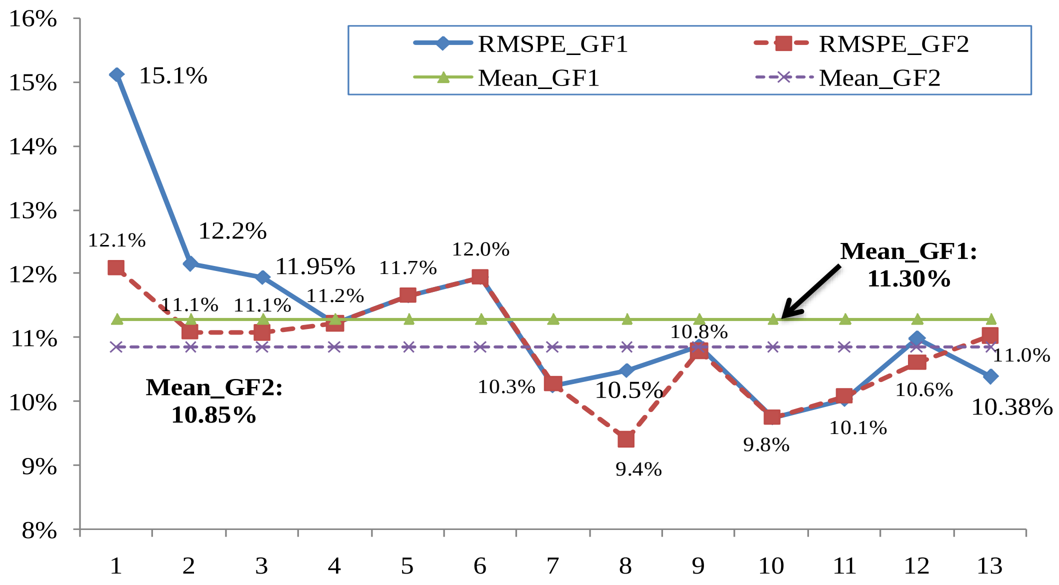
Country /Region	*Quarter*	APE_{SF} (a)	APE_{GF1} (b)	APE_{GF2} (c)	$\%_{(b-a<0)}$	$\%_{(c-a<0)}$	$\%_{(c-b<0)}$
Taiwan	2010Q1	14.54	12.19	13.53	83	100	50
	2010Q2	14.45	11.79	11.16			
	2010Q3	7.36	4.94	5.83			
	2010Q4	0.63	1.60	0.54			
	2011Q1	9.18	6.69	8.08			
	2011Q2	18.30	15.58	14.98			
Japan	2010Q1	5.71	6.05	5.38	50	83	50
	2010Q2	24.33	21.79	21.95			
	2010Q3	3.43	3.64	3.16			
	2010Q4	2.18	1.83	1.97			
	2011Q1	2.32	3.02	2.53			
	2011Q2	12.74	11.12	12.22			
Australia	2010Q1	0.59	1.96	3.93	33	33	17
	2010Q2	0.45	2.11	4.05			
	2010Q3	10.51	8.28	6.98			
	2010Q4	2.23	0.34	1.92			
	2011Q1	3.58	5.37	8.04			
	2011Q2	0.29	1.72	4.37			
UK	2010Q1	0.43	4.23	3.09	0	0	100
	2010Q2	11.13	15.41	14.23			
	2010Q3	6.41	10.44	9.16			
	2010Q4	2.79	6.90	5.66			
	2011Q1	11.42	15.24	14.23			
	2011Q2	13.33	16.99	16.42			
USA	2010Q1	6.47	4.21	4.40	100	100	67
	2010Q2	8.56	5.70	4.96			
	2010Q3	5.80	3.18	2.60			
	2010Q4	5.07	2.37	2.30			
	2011Q1	8.52	6.30	6.53			
	2011Q2	9.09	6.36	5.52			
Mean		**8.86**	**8.02**	**7.99**	**61**	**69**	**61**

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

Figure 4.6 shows the forecasting performances of 13 of the individual participants involved in the survey for both rounds. Seven of the 13 produced out-of-sample forecasts that were better than the statistical forecasts (according to both the MAPE and RMSPE measures). In order to test whether the overall group performance improved over the rounds, the differences were analysed using a paired *t*-test. The judgmental group forecasts in the second round were shown to be significantly more accurate than those in the first round at the 10% significance level, as measured by the MAPE (t (12) = –1.418, p = 0.091). This was further confirmed to be the case by the RMSPE value (t (12) = –1.737, p = 0.054). Thus, the group performance improved significantly with the use of the Delphi approach. This finding is consistent with the results of previous studies, reporting that incorporating forecasters' technical and contextual knowledge into the statistical forecasts helps to improve the forecast accuracy (Sanders & Ritzman, 1995). The participants in this study were postgraduate (mainly PhD) students and research assistants in tourism and hospitality management with a certain degree of domain knowledge about the development and prospects of the Hong Kong tourism industry. The effects of several special events that occurred during the forecast period were not considered in the statistical models but were incorporated as judgmental inputs, including the recovery from the 2008 financial crisis, the floods in Australia, the earthquake in Japan, the construction of the high-speed railway between Hong Kong and Mainland China, and the London Olympic Games slated for 2012. In addition, the majority of the participants (76%) were well-trained in quantitative methods, which helped them to understand the statistical forecasting procedure.



(a) MAPE



(b) RMSPE

Figure 4.6 Individual participants' forecasting performances over rounds

4.4 A Feedback Survey

A feedback survey was distributed to all of the participants to examine how they perceived the effectiveness of the HKTDFS. The questionnaire (see Appendix E) was structured carefully using a 5-point Likert scale, ranging from 1 ("*strongly* *disagree*”) to 5 (“*strongly agree*”) for all statements. Among the 12 positive responses, over half (58%) of the participants agreed that the “forecasting system is easy to use”, whereas the rest had no strong opinion. About 25 per cent of the participants strongly agreed that they had a clear understanding of what they were expected to do in the forecasting tasks after reading the instructions. About 42 per cent “agreed” (albeit not strongly), and only 8 per cent felt that they were unclear about the tasks even when they were provided with a step-by-step video demonstration in addition to the written instructions. Regarding the time needed to complete the survey, 67 per cent of the participants thought that the time allotted (ranging from 20 to 40 minutes) was appropriate. When asked to evaluate the statistical feedback from the first round of the survey, half of the participants agreed that it was useful for assisting their adjustments in the second round. The participants indicated that the tabular and graphical data summaries were useful: 17 per cent “strongly agreed” and 75 per cent “agreed” that the graphical presentation was useful, whereas 8% “strongly agreed” and 67 per cent “agreed” that the tabular information was useful. The majority of the respondents (84%) agreed that the “graphs on the website are more informative than the tables”. Regarding the amount of historical data, approximately 60 per cent of the participants agreed that the current system provided sufficient data to assist them with the adjustments. The participants were also asked to provide suggestions on the amount of historical data needed. 42 per cent suggested data covering the previous five years, 42 per cent suggested 10 years, and the remaining 16 per cent suggested periods ranging from less than 5 years to more than 10 years.

4.5 Chapter Summary

The Web-based HKTDFS is proposed as an innovative online platform for tourism demand forecasting that takes full advantage of web technologies and advanced tourism demand forecasting techniques. Like other Web-based systems, the HKTDFS has four main attributes: wide accessibility, flexibility, reusability, and user-friendliness. In addition, various new features distinguish the current HKTDFS from other forecasting support systems, including: (a) integrating statistical and judgmental forecasts through a dynamic online Delphi survey; (b) creating different scenarios based on user-customized specifications; and (c) applying JSP, which provides a connection to the R Engine. One significant benefit to tourism practitioners is that the HKTDFS allows the integration of quantitative and judgmental forecasts in a Web-based forecasting system.

Overall, the experimental study showed that a greater forecast accuracy was achieved with the judgmentally adjusted statistical forecasts than with the statistical forecasts alone. In addition, the integration of statistical and judgmental forecasts improved the forecast accuracy for four of the six source markets of interest (i.e. China, Taiwan, Japan, and the USA). The long-haul markets tended to produce more accurate forecasts than the short-haul markets; however, more remarkable improvements were found for the three short-haul markets. This is probably due to the relatively more stable data patterns for the arrivals data of the long-haul markets, and the high capacity of the econometric models, which produce very accurate forecasts. Thus, including judgmental inputs for such series did not significantly improve the forecast accuracy; on the contrary, it harmed the forecast accuracy. The benefits of including judgmental inputs in quantitative forecasts depend on the characteristics of the data series being examined. These results suggested that the accuracy of the judgmental forecasts increased in the second round relative to the first round, with a reduction of the MAPE from 8.02 to 7.99 per cent. The paired *t*-test results confirmed that there is a significant reduction in the MAPE and RMSPE over two rounds.

The forecasting performance of the HKTDFS is achieved through the following factors. First, an advanced econometric modelling method (i.e. the ARDL-ECM) is used to estimate the demand model for each source market. Second, the HKTDFS provides flexible adjustment options for the forecasters to adjust their forecasts by either the year or the quarter over different forecasting periods. Third, the system provides useful feedback about the summary forecasts generated by all of the experts in the early rounds with both high-resolution graphs and tables, so that the Delphi experts are well informed for subsequent adjustments. Fourth, the use of a Web-based platform allows users to access the system anytime and anywhere and allows collaboration between individuals in different geographic locations and representing different knowledge domains. Finally, participants who have a high level of technical knowledge of tourism demand forecasting, as well as some degree of contextual knowledge of the Hong Kong tourism industry, may contribute to the improvement in accuracy. This is supported by Sanders and Ritzman's (1995) finding that the combination of statistical forecasts and judgmental forecasts based on contextual knowledge could lead to significantly more accurate forecasts. It is worth noting that this study did not distinguish the factors that influence the accuracy of judgmental inputs into the econometric model or the forecasting system. A number of factors, such as the form of data presentation (with tables and graphs), the provision of feedback over different rounds of the Delphi survey, the capability of good functional forecasting modules, clear instructions for the Delphi survey, and the inclusion of historical data series, can lead to a good forecasting performance. Further research is thus needed to investigate which attributes contribute most to the improvement in accuracy.

To sum up, the main purposes of the pilot study are to (a) test the reliability of the Judgmental Forecasting Module, (b) evaluate the forecasting performance of the proposed integrated framework using postgraduate students and research staff as the respondents, and (c) investigate participants' perception regarding the effectiveness of the HKTDFS. The findings of the pilot study provided preliminary evidence to validate the forecasting ability of the proposed integration framework by using a Web-based forecasting system (HKTDFS) and utilizing a structured group forecasting technique (Delphi) to quantify and elicit experts' opinions. The feedback comments obtained from the student participants helped to check the appropriateness of the Delphi survey (e.g. length of completing the survey, contents of instructions, presentation format of data, use of information provided, etc.) as well as to further enhance the functional ability of the forecasting system. In addition, to reflect the "true" forecasting ability of the HKTDFS, experts from the tourism industry are regarded as more qualified Delphi panellists than student participants.

Chapter 5 : Findings and Discussions

5.1 Introduction

To recap, Chapter 3 provided details about how to model and forecast tourism demand in Hong Kong, apply the judgmental adjustment procedure via the Delphi approach in the HKTDFS, and evaluate the forecasting performance. The methods described in Chapter 3 help to guide the data analysis in this chapter.

The preceding chapter showed the results from a pilot study using postgraduate students as Delphi participants which suggested the validity of integrating judgmental inputs into statistical forecasts. The pilot experiment served as the pretesting stage before conducting the main Delphi survey. Before launching the main survey, a few revisions were made to refine the Judgmental Adjustment module in the HKTDFS according to student participants' feedback and comments; for example, statistical forecasts were updated based on the 1985Q1−2010Q4 sample.

The main purpose of this chapter is to test the research hypotheses proposed in Chapter 3. The remaining sections go into detail on such aspects as the validity and reliability of the econometric modelling and forecasting results, the main Delphi forecasting results, the forecasting performance of the statistical and judgmental forecasts, and the in-depth interview results.

This chapter starts with two sections on the econometric analysis of tourism demand: the ARDL bounds test results and the diagnostic test results of the econometric (or statistical) models. A brief recap of the Delphi survey procedure and the main forecasting results (forecasts and comments from experts) from the six source markets are presented in Sections 5.3 and 5.4. Section 5.5 − the key section presenting the results of the hypothesis testing − provides an extensive analysis of the efficiency, biasness, and accuracy of the statistical and judgmental forecasts. In addition to the evidence from the existing literature, in-depth interviews provide the experts' perspective to further explore the possible reasons for improvements or deteriorations in accuracy with regard to the integrative forecasting approach. The results from the in-depth interviews are presented in Section 5.6. As before, a brief summary of the whole chapter is provided at the end.

5.2 Econometric Analysis of Tourism Demand

This section presents the results of the unit root tests, the ARDL bounds tests, and the diagnostic tests. The unit root tests were used to ascertain the order of integration and to make sure that all of the variables included in the ARDL-ECM models were either $I(0)$ or $I(1)$. The ARDL bounds tests were followed up to examine the existence of short- and long-run relationships between visitor arrivals and the key determinants. Diagnostic tests were employed to check the adequacy of the estimated models before using them to make statistical forecasts.

5.2.1 Unit root test results

The analysis began by investigating the unit root test of variables. Table 5.1 presents the nonstationarity and stationarity test results for all of the dependent and independent variables in level and first differences using ADF, KPSS, and NP tests. As shown in Table 5.1, the null hypothesis of a unit root was not rejected for all variables in the ADF test; it was only rejected in two cases − visitor arrivals in the Japan model and own price in the USA model. Taking first differences rendered all series stationary, with the ADF statistics in all cases (except for the two mentioned above) being less than the critical values at either the 1% or 5% significance level. The KPSS and Ng-Perron test results also showed that all of the variables concerned were $I(1)$. As the results showed that all of the variables in the models were either $I(0)$ or $I(1)$, the ARDL bounds test was valid.

Table 5.1 Unit root test results

Variable	Level				First difference				
	ADF	KPSS	NP(MZ_{α}^{d})	NP(MZ_{t}^{d})	ADF	KPSS	NP(MZ_{α}^{d})	NP(MZ_{t}^{d})	
$VA_{Australia}$	-2.99	0.20*	-1.89	-0.63	-9.07**	0.17	-48.97**	-4.95**	$I(1)$
VA_{China}	-2.74	1.14**	-13.61	-2.59	-5.64**	0.06	-46.72**	-4.83**	$I(1)$
VA_{Japan}	-4.90**	0.18	-26.58**	-3.64**					$I(0)$
VA_{Taiwan}	-2.92	0.24**	-3.09	-1.13	-7.77**	0.36	-49.37**	-4.96**	$I(1)$
VA_{UK}	-2.73	1.83**	-6.49	-1.80	-12.45**	0.00	-277.95**	-11.78**	$I(1)$
VA_{USA}	-1.91	1.03**	-0.90	-0.40	-10.35**	0.03	-44.78**	-4.72**	$I(1)$
$Y_{Australia}$	-1.45	0.17*	-4.39	-1.44	-9.92**	0.13	-25.22**	-3.53**	$I(1)$
Y_{China}	-2.70	0.19*	-2.88	-1.02	-3.22*	0.11	-28.60**	-3.78**	$I(1)$
Y_{Japan}	-2.27	0.23*	-3.17	-1.10	-4.06**	0.09	-21.07*	-3.24*	$I(1)$
Y_{Taiwan}	-1.74	0.30**	-2.53	-1.04	-2.99*	0.36	-11.39*	-2.38*	$I(1)$
Y_{UK}	-0.71	5.64**	-5.24	-1.36	-5.31**	0.24	-26.09**	-3.49**	$I(1)$
Y_{USA}	-1.37	1.14**	-10.15	-2.04	-4.32**	0.31	-22.34**	-3.34**	$I(1)$
$P_{Australia}$	-1.34	0.28**	-4.08	-1.27	-7.30**	0.43	-22.69**	-3.36**	$I(1)$
P_{China}	-1.96	0.28**	-0.83	-0.61	-4.46**	0.42	-32.79**	-4.05**	$I(1)$
P_{Japan}	-2.51	0.50*	-10.68	-2.29	-4.40**	0.11	-16.66**	-2.89**	$I(1)$
P_{Taiwan}	-1.19	0.78**	-1.27	-0.70	-6.76**	0.15	-32.78**	-4.01**	$I(1)$
P_{UK}	-1.04	4.78**	-4.38	-1.17	-9.83**	0.29	-28.68**	-3.76**	$I(1)$
P_{USA}	-2.97**	0.33	-10.43*	-2.27*					$I(0)$
$PS_{Australia}$	-1.46	0.16*	-4.85	-1.52	-7.01**	0.15	-44.78**	-4.68**	$I(1)$
PS_{China}	-1.96	0.56*	-13.67	-2.61	-11.36**	0.03	-49.92**	-5.00**	$I(1)$
PS_{Japan}	-2.13	0.92**	-8.22	-2.02	-6.28**	0.11	-41.03**	-4.50**	$I(1)$
PS_{Taiwan}	-2.51	0.20*	-5.60	-1.65	-9.81**	0.14	-50.18**	-5.00**	$I(1)$
PS_{UK}	-1.42	8.16**	-5.64	-1.61	-7.00**	0.28	-43.37**	-4.61**	$I(1)$
PS_{USA}	-1.77	0.98**	-6.21	-1.73	-7.07**	0.12	-45.12**	-4.72**	$I(1)$

Note: (1) ** and * denote rejection of the null hypothesis based on the critical values from MacKinnon (1996), Kwiatkowski et al. (1992), and Ng and Perron (2001), at the 1% and 5% significance level, respectively. (2) For the ADF and NP tests, the null hypothesis is that there is a unit root in the test series; whereas for the KPSS test, the null hypothesis is that there is no unit root in the test series. The optimal lag of respective model is determined either by AIC or SC.

5.2.2 ARDL bound test results

The unit root test results presented in Section 5.2.1 show that all of the variables were integrated of $I(0)$ or $I(1)$. This suggests that it was appropriate to use the bounds testing procedure. The first stage of the ARDL cointegration method involved comparing the calculated *F*-statistics with the critical values for testing the null hypothesis of a joint significance test that implied no cointegration. As presented in Panel A of Table 5.2, the calculated *F*-statistics exceeded the upper bound critical value at least at the 10% level for all six markets. This implied that the null hypothesis of no cointegration cannot be accepted. The results from the application of the bounds *t*-test to the six models (except for China) clearly rejected the null hypothesis, suggesting the existence of a long-run relationship among income, own price, substitute price, and the lagged dependent variable.

Taking the demand model for Australian tourists as an example, the relevant *F*-statistic was 43.79, which was greater than the upper critical bound value 7.84 at the 1% level. The null hypothesis (H_0: $\pi_1 = \pi_2 = \pi_3 = 0$) was conclusively rejected. For the bounds *t*-test, the model rejected the null as the *t*-statistic was -12.99, exceeding the upper critical bound value of -3.82 at the 1% level. The *F*-test and *t*-test results indicated the existence of cointegration in the Australia model. The only model that did not conclusively reject the null of the bounds *t*-test was the China model; therefore, caution should be taken when interpreting its modelling and forecasting results.

Table 5.2 Diagnostic test and bounds test results

Test statistic	Australia	China	Japan	Taiwan	UK	USA
Panel A: Bound tests						
F statistic	43.79***	4.83*	4.89**	13.11***	75.97***	13.66***
t statistic	-12.99***	-2.58	-3.99**	-6.16***	-17.08***	-7.11***
Lag	1	3	3	2	1	1
Panel B: Model fitting						
R^2	0.86	0.77	0.94	0.84	0.94	0.95
Adjusted R^2	0.84	0.73	0.93	0.82	0.93	0.94
F statistic	54.53***	19.15***	77.14***	31.35***	107.44***	107.31***
DW statistic	1.54	2.12	2.22	1.82	1.58	1.94
AIC	-1.91	-2.17	-2.15	-1.61	-1.90	-2.41
Panel C: Diagnostic tests						
J-B test	22.24***	2.64	4.60	4.34	1.89	46.00**
LM test	5.96**	6.13	3.88	5.64	3.06	3.20
Breusch-Pagan-Godfrey test	10.89	20.78	18.03	40.89**	23.49**	22.92
White test	13.68	28.63**	14.99	43.23***	16.15	28.10**
ARCH test	3.18	0.02	0.52	5.46**	0.27	0.23
RESET	10.40***	1.30	2.18	11.35***	8.02**	8.69***
Panel D: Demand elasticities						
Income	1.14***	1.81***	2.17*	0.33	1.44***	1.16***
Own price	-0.38***	-0.81*	-1.03**	-1.57	-0.10*	-0.25*
Cross price	0.07	-1.35	-0.56	0.34	0.67***	-0.13

Note: ***, ** and * indicate significance at the 1%, 5%, and 10% level, respectively.

5.2.3 Diagnostic test results

In order to ensure an appropriate model, a set of diagnostic tests, each of which was designed to detect a particular form of model inadequacy (e.g. autocorrelation, heteroscedasticity, nonnormality, model misspecification, etc.), was carried out for the six final models that were estimated by Equation (3.4).

Table 5.2 (Panel B) shows that all six models had a high goodness of fit, as suggested by the high values of the adjusted R^2: 0.73 (China), 0.84 (Australia), 0.82 (Taiwan), 0.93 (Japan), 0.93 (UK), and 0.94 (USA). This also means that about 73% (China), 84% (Australia), 82% (Taiwan), 93% (Japan), 93% (UK), and 94% (USA) of the variations in visitor arrivals from the relevant markets over the period 1985Q1−2010Q4 could be explained by the regressors of the ARDL-ECM models.

The diagnostic statistics (Panel C in Table 5.2) show that the Japan model passed all of the tests while the other five models passed most of the five tests but failed some of them. Four models (China, Japan, Taiwan and the UK) passed the J-B normality test, while the other two models (Australia and the USA) failed the normality test. Despite the violation of the normality test for the Australia and USA models, the Gauss-Markov theorem shows that the OLS estimators are still the best linear unbiased estimators (BLUE) (Gujarati & Porter, 2002) under the other assumptions (e.g. homoscedasticity, no autocorrelation between the disturbances, no perfect multicollinearity, etc.).

The DW statistics for all markets fell within the acceptable range of 1.50 to 2.50. There was no evidence of autocorrelation in the disturbance of the error term for five models (China, Japan, Taiwan, the UK, and the USA) according to the LM tests, which satisfied the assumption of the independence of errors. The Australia model failed the autocorrelation test; however, this is often the problem when a lagged dependent variable is included as an explanatory variable in a model because the explanatory variable would probably be highly correlated with the lagged demand (Morley, 2009).

Four models − Australia, China, Japan, and the USA − were free of the heteroscedasticity problem according to the B-P test; however, the White tests (without White cross terms) suggested that three models (Australia, Japan, and the UK) did not have a heteroscedasticity problem. The ARCH tests suggested that the errors were homoscedastic and independent of the regressors in five of the six models (except for the Taiwan model). Four models, namely Australia, Taiwan, the UK, and the USA, failed the RESET test for model misspecification. Overall, these six models were valid due to the satisfactory diagnostic testing results and were reliable for further analysis.

5.2.4 Tourism demand elasticities

The empirical results of the long-run tourism demand model for Hong Kong's six key tourist source markets, which were obtained by normalizing on visitor arrivals, are presented in Table 5.2. As expected, the signs for the income variables were positive for all six markets. Income elasticities were significant at least at the 10% level for five out of the six models, suggesting that the incomes of origin countries/regions are the key influencing factor in determining visitor arrivals to Hong Kong, but the magnitudes of the estimated elasticities varied across markets. Hong Kong is likely to gain as real income in the origin markets rises– the results in Table 5.2 indicate that a 1% increase in income will lead to a 1.14%, 1.81%, 2.17%, 0.33%, 1.44%, and 1.16% increase in visitor arrivals from Australia, China, Japan, Taiwan, the UK, and the USA, respectively. The point estimates of income elasticity for five of the six models were greater than one, suggesting that travelling to Hong Kong is generally regarded as a luxury product by visitors from these source markets. The income elasticity for Taiwan was less than one, suggesting that the demand for Hong Kong tourism from Taiwanese visitors is income inelastic. This finding is consistent with Chon et al.'s (2010) study; they explained that the plausible reason for this finding is that "a high proportion of Taiwanese visitors in Hong Kong are transit passengers, who regard Hong Kong as the gateway to and from Mainland China or other destinations" (p. 268).

In line with the law of demand, almost all of the estimated own-price elasticities were negative, the only exception being China. A negative value of own-price elasticity indicates that an increase in the price of tourism goods/services in Hong Kong would lead to a decline in the demand for Hong Kong tourism. It was also found that the point estimate of own-price elasticity for three long-haul markets (Australia, the UK, and the USA) was significantly less than one, revealing that visitors from these three countries are relatively less sensitive to the price changes of tourism products/services in Hong Kong.

Compared to the income and own-price variables, the impacts of the substitute price variables were found to be less influential on tourism demand in Hong Kong as the substitute price was found to be statistically significant for only one model (UK). The estimated cross-price elasticities were positive in the Australia (0.07), Taiwan (0.34), and UK (0.67) models, which means that an increase in the costs of tourism in competing destinations would lead to an increase in the demand for Hong Kong tourism. This suggests that tourists from these three markets are very much aware of the costs of tourism in the six destinations that Hong Kong competes with (i.e. Mainland China, South Korea, Japan, Singapore, Thailand, and Taiwan), and that a change in the cost of holiday travel in these competing destinations will have a big impact on the demand for Hong Kong tourism from visitors from the above three markets. Therefore, there is a need to launch a more aggressive destination image branding campaign while reinforcing already held images to tailor experiences to this target audience. The cross-price elasticities of the China (-1.35), Japan (-0.56) and USA (-0.13) models were found to be negative.

5.3 Basic Information About the Main Delphi Survey

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. Between Round 1 and Round 2, three of the 18 panellists failed to respond to the survey; the addition of two new experts led to a total of 17 experts completing Round 2. In total, 15 experts took both the first and second round surveys.

All of the panellists were asked to self-rate their level of expertise in tourism demand forecasting on a 7-point Likert scale. The mean self-rating scores for the first and second rounds of the Delphi survey were 4.83 and 4.53, respectively (see Table 5.3). Of these 18 participants, 11.1 per cent rated themselves as having very little expertise; 5.6 per cent rated themselves as having level-7 experience; about 66.7 per cent rated themselves at levels 5 and 6; and the remaining 16.7 per cent fell between levels 3 and 4.

Table 5.3 Self-rating of expertise over rounds

Self-rating expertise	R1		R2	
	Count	%	Count	%
1	2	11.1	***3***	17.6
3	1	5.6	1	5.9
4	2	11.1	2	11.8
5	5	27.8	***4***	23.5
6	7	38.9	***6***	35.3
7	1	5.6	1	5.9
Industry	7		6	
Academic	11		11	
Total	18	100.0	17	100.0
Industry	4.57		4.50	
Academic	4.91		4.45	
Mean rating	4.83		4.53	

The results revealed a convergence of group opinion and an increase in convergence between rounds. Table 5.4 indicates that the reductions in standard deviation value are noteworthy: the standard deviations for the second round were all smaller than the first round, indicating that consensus was well reached in the second round. Thus, it was decided that consensus had been reached and that further rounds would not produce any additional convergence of opinion.

Table 5.4 Standard deviations for six source markets over rounds

Quarter	Australia			China			Japan			Taiwan			UK			USA		
	R1	R2	Gap (%)	R1	R2	Gap (%)	R1	R2	Gap (%)	R1	R2	Gap (%)	R1	R2	Gap (%)	R1	R2	Gap (%)
2011Q2	5.8	5.0	-13.1	562.4	303.5	-46.0	54.3	50.5	-7.0	14.1	7.8	-44.7	2.8	0.6	-79.8	11.0	3.1	-71.3
2011Q3	6.1	4.0	-34.3	679.9	325.1	-52.2	57.2	53.7	-6.0	16.8	9.5	-43.5	2.7	0.5	-79.8	10.2	3.6	-64.9
2011Q4	6.9	3.6	-47.1	718.5	349.7	-51.3	56.2	53.9	-4.2	15.3	8.7	-43.2	3.4	0.5	-84.6	11.9	4.2	-64.9
2012Q1	5.8	2.7	-54.1	733.6	358.2	-51.2	13.1	9.5	-27.6	16.2	9.1	-43.8	4.8	2.0	-57.5	10.2	3.9	-62.2
2012Q2	6.0	3.1	-48.7	669.2	337.3	-49.6	12.1	8.8	-27.6	15.1	8.8	-42.2	3.3	1.5	-54.1	11.2	3.9	-64.9
2012Q3	7.1	4.4	-38.2	760.2	361.3	-52.5	12.9	9.4	-27.6	17.0	8.5	-49.8	19.7	19.3	-2.5	10.5	3.7	-64.9
2012Q4	7.1	3.8	-47.1	810.4	363.7	-55.1	13.0	9.4	-27.6	15.5	7.9	-49.2	4.0	1.6	-59.2	12.2	4.3	-64.9
2013Q1	6.0	2.8	-54.1	826.8	398.8	-51.8	13.6	9.8	-27.6	16.7	8.4	-50.0	4.6	1.4	-69.7	10.5	4.0	-62.2
2013Q2	6.2	3.2	-48.7	738.8	378.8	-48.7	12.6	9.1	-27.6	15.8	8.0	-49.3	2.9	0.6	-79.8	10.1	3.3	-67.4
2013Q3	6.6	3.8	-41.8	842.6	417.2	-50.5	13.4	9.7	-27.6	17.7	8.6	-51.1	2.8	0.6	-79.8	9.5	3.1	-67.4
2013Q4	7.4	3.9	-47.1	902.3	429.4	-52.4	13.4	9.7	-27.6	16.1	8.0	-50.5	3.6	0.6	-84.6	11.1	3.6	-67.4
2014Q1	6.2	2.9	-54.1	928.7	397.6	-57.2	14.0	10.1	-27.6	17.0	8.5	-50.0	3.9	0.7	-81.6	9.5	3.4	-64.1
2014Q2	6.4	3.3	-48.7	829.5	428.3	-48.4	13.0	9.4	-27.6	16.0	8.1	-49.3	3.0	0.6	-79.7	10.4	3.4	-67.4
2014Q3	6.8	4.0	-41.8	955.6	418.8	-56.2	13.8	10.0	-27.6	17.1	8.8	-48.7	2.9	0.6	-79.7	9.7	3.2	-67.4
2014Q4	7.6	4.0	-47.1	1029.7	487.6	-52.7	13.8	10.0	-27.6	15.6	8.1	-48.0	3.7	0.6	-84.4	11.4	3.7	-67.4
2015Q1	6.4	2.9	-54.1	1058.3	453.1	-57.2	14.4	10.4	-27.6	16.7	8.6	-48.4	4.0	0.7	-81.6	9.7	3.5	-64.1
2015Q2	6.6	3.4	-48.7	975.2	545.9	-44.0	13.7	9.6	-29.8	15.5	8.2	-46.6	3.1	0.6	-79.7	10.7	3.5	-67.4
2015Q3	7.0	4.1	-41.8	1120.2	541.1	-51.7	14.6	10.2	-29.8	17.4	8.9	-48.7	3.0	0.6	-79.7	10.0	3.3	-67.4
2015Q4	7.8	4.1	-47.1	1206.3	693.2	-42.5	14.6	10.2	-29.8	15.9	8.2	-48.0	3.8	0.6	-84.4	11.7	3.8	-67.4
Min			-54.1			-57.2			-29.8			-51.1			-84.6			-71.3
Max			-13.1			-42.5			-4.2			-42.2			-2.5			-62.2

5.4 Main Findings of the Delphi Survey

This section summarizes the key main findings from the two-round Delphi survey. To start with, a brief summary is provided below to depict the future projections of visitor arrivals over 2011–2015 for the six tourist origin markets of Hong Kong.

Strong economic growth and a closer relationship between Hong Kong and the Mainland are likely to continue to fuel the inbound demand from Mainland visitors. The experts predicted an annual increase of 11.03 per cent in visitor arrivals from China over 2010–2015. However, some experts asserted that the group forecasts might have overestimated the growth figures for China. It was projected that the Taiwan market, the second largest source market for Hong Kong, would be affected by the global economic crisis and the availability of direct flights between the Mainland and Taiwan. As a result, predicted arrivals to Hong Kong from Taiwan were not expected to exceed the pre-crisis level until 2014. As anticipated by most experts, inbound tourism from Japan would be significantly affected by the Japanese earthquake in 2011. It was also believed that the gloomy economic prospects in Japan would likely drag down growth in arrivals. The experts thus foresaw a negative growth rate of 1.21 per cent in 2011 compared to 2010.

The panellists expected the poor economic outlook to continue to affect tourism demand from the USA. It was predicted that the number of US tourists visiting Hong Kong would grow by around 2.38 per cent per annum from 2010 onwards, reaching 1.32 million in 2015. In view of the optimistic economic development prospects of Australia and its strong currency, most of the panellists believed that tourism demand from Australia would recover quickly from the 2009 setback and continue to grow over the forecasting period. Due to the influence of the London Olympic Games, some experts believed that inbound demand from the UK would suffer. However, the group forecasts suggested an increase of 6.2 per cent in arrivals from the UK in 2012.

5.4.1 Arrivals from Australia

Most of the experts believed that visitor arrivals from Australia would grow from 2011 onwards due to the strong Australian currency and optimistic economic development prospects (see Table 5.5). It was forecast that visitor arrivals from Australia would reach 0.65 million by 2011, a slight increase of 0.45 per cent compared to 2010. This figure was expected to reach 0.75 million by 2015 (see Figure 5.1), representing a minor annual increase of 2.9 per cent from 2010. However, one expert believed that "the statistical forecasts of Australian visitor arrivals to Hong Kong were too high" and expressed his concern as to whether economic growth would be sustained at the current level.

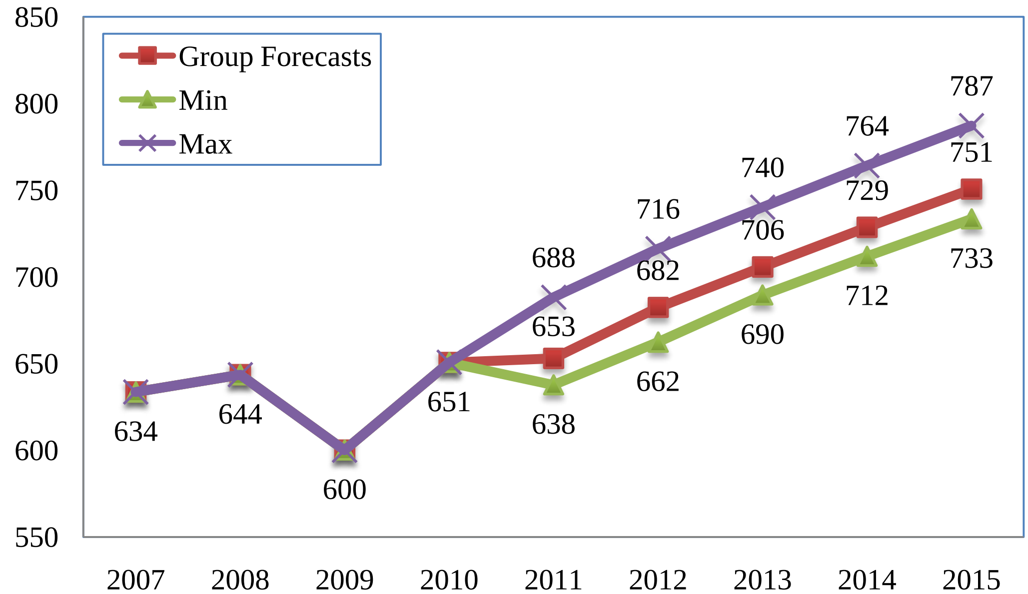


Figure 5.1 Annual forecasts of visitor arrivals from Australia ('000), 2007−2015

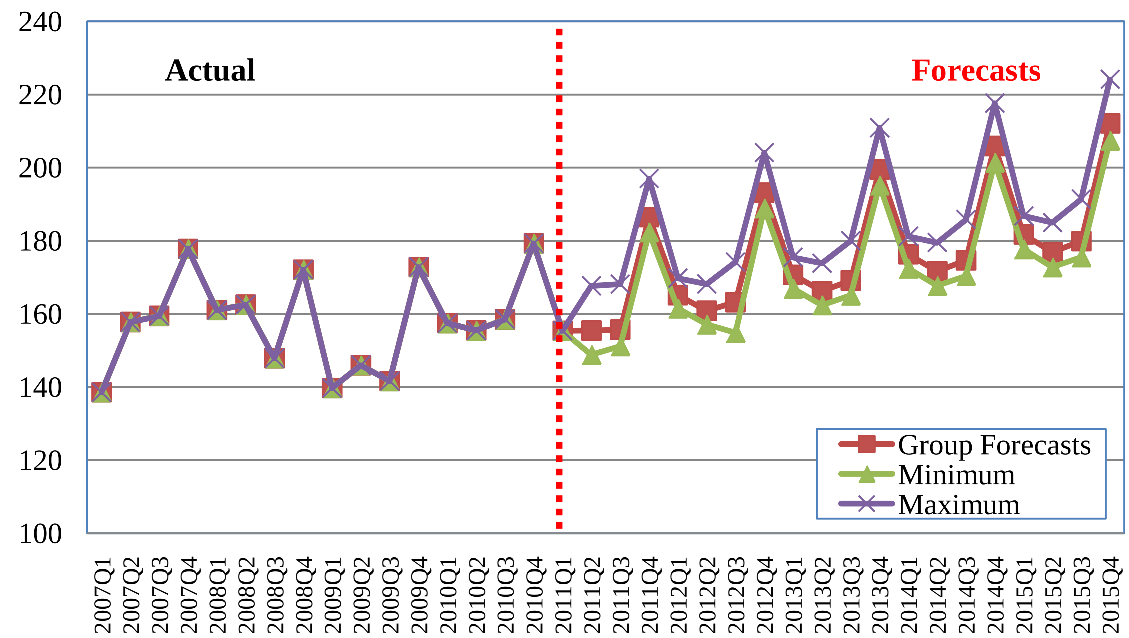


Figure 5.2 Quarterly forecasts of visitor arrivals from Australia (‘000), 2007Q1−2015Q

Table 5.5 Comments for Australia

Projections		**Comments**
Optimistic	R1	1. The **strong Australian currency**.
		2. **Strong Australian dollars** and stronger relations/trade between Australia and China.
		3. Australia’s overall **economic growth** will be **strong** in the next decade given its link to China.
		4. I am **confident** with the **economic development** of Australia in the near future.
		5. With the **strong Australian dollar** and the **increasing air capacity**, this market will definitely **improve**. I predicted an 8% growth in 2012 and 5% year over year thereafter.
		6. Australian arrivals have been quite **volatile** over the past decades. Whether the **strong AUD against HKD** could be sustained in the coming few years **is hard to predict**. Yet the strong Australian outbound market provides **solid support** for arrivals growth to Hong Kong.
	R2	1. No adjustment is made. I still believe that Australian market is strong and the trend for arrivals will continue to grow.
Pessimistic	R1	1. The **economic growth** of Australia **would not sustain** at the current level.
	R2	2. I stick to my previous adjustments, as the statistical forecasts of Australian tourist arrivals to Hong Kong were too high.

5.4.2 Arrivals from China

Over the past few years, visitors from China have formed the lion's share of total visitor arrivals in Hong Kong. The majority of the experts agreed that visitor arrivals from the Mainland would increase continuously in the next 5 years. Strong economic growth in China and a closer relationship between Hong Kong and Mainland China were the two most frequently mentioned driving forces for Hong Kong's tourism industry over the forecasting period. The number of Mainland Chinese visitors to Hong Kong in 2011 was estimated to reach 25.04 million, representing an annual increase of 11.69 per cent (see Figure 5.3). Visitor arrivals from Mainland China in 2015 were predicted to reach 37.66 million, representing an annual increase of 11.03 per cent from 2010. Quarterly adjustments were made particularly for the second quarter of 2011−2015 due to the impact of the Ching Ming Festival, Labour Day, and Dragon Boat Festival holidays (see Table 5.6). One expert stated that "there does not seem to be any obvious reason for the growth of this market to slow down too sharply in the near future (Q2 and Q3 of 2011)"; he also believed the future growth of Chinese market would slow down a bit and that "seasonality should not be so obvious (Q2 arrivals appear relatively lower compared to other seasons)". Another expert agreed that the second quarter arrivals forecasts should be upwardly adjusted.

Four experts held relatively pessimistic opinions about the future growth of tourism demand from the Mainland because they thought that China's economic growth would slow down in the near future. One expert from the academic arena stated that "[the] majority of the experts might have overestimated the growth for China" and explained that "[t]he China market will grow; however, as keen competition increases, this market will grow at a slower pace". In addition, the availability of more substitute destinations will lead to more competition for Hong Kong.

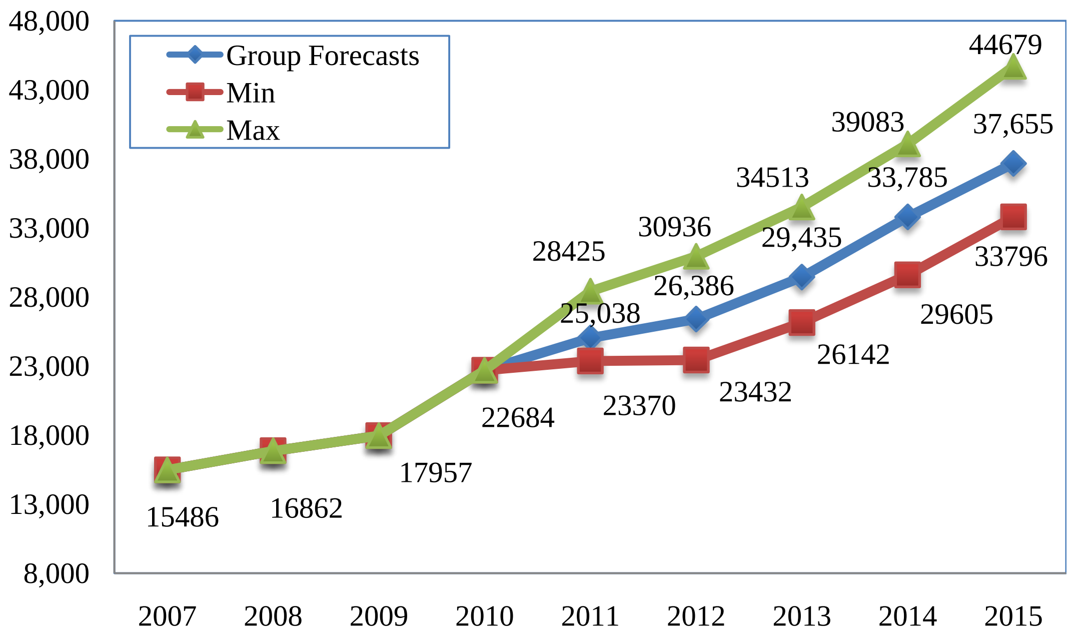


Figure 5.3 Annual forecasts of visitor arrivals from China ('000), 2007−2015

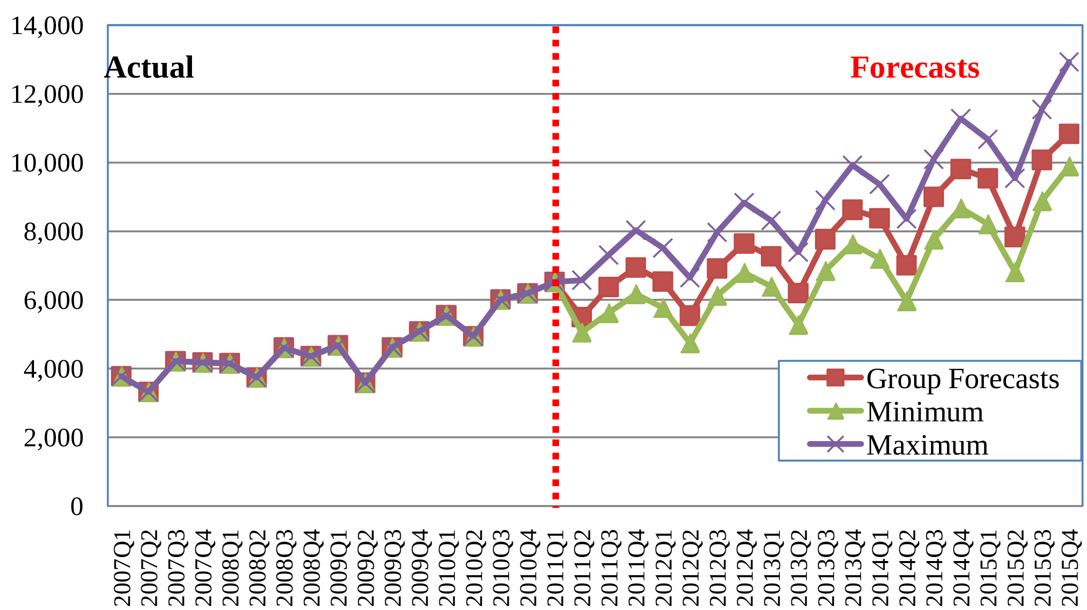


Figure 5.4 Quarterly forecasts of visitor arrivals from China ('000), 2007Q1−2015Q4

Table 5.6 Comments for China

Projections		Comments
Optimistic	R1	1. Visitor arrivals from Mainland China will **increase continuously** due to favourable foreign exchange rates, convenient transportation links, and increasing business activities between Hong Kong and China.
		2. The **growth** in the **Chinese economy** will continue and the number of people entering the middle-income bracket will grow. So, on the **demand** side, demand will be **increasing,** especially with increasingly easy access between Hong Kong and the Mainland. On the **supply** side, the number of beds available in Hong Kong will increase, and the standard of service will be improved gradually. Some of the existing properties will be adjusted to cater for the increase in demand. **Macau** is the **direct competitor**. However, learning from the example of Macau, Hong Kong could develop **casinos** to gain the potential benefits for the people of Hong Kong.
		3. I believe that there will be a **continuous increase** in Chinese tourists visiting Hong Kong because the relationship between Hong Kong and China will be even **more connected** in the future and **China's economy** will **continue to grow rapidly**.
		4. The China market is the most important market for Hong Kong due to the large number of arrivals. This market will continue to grow as per capita income increases in China. The recent tax cuts will help the middle class and increase their spending power. However, the opening of free and easy travel to Taiwan and Japan will definitely dilute the arrivals to Hong Kong and affect the growth momentum.
		5. China's **continuing strong economic** performance and the new **high speed railway** will **increase accessibility** between Hong Kong and China.
		6. [There will be] more visitors from the Mainland.
		7. The statistical model projected a drop in Mainland arrivals in 2011 but picked up the rising trend at an increasing rate in the later years. The drop in arrivals was probably caused by the impact of **H1N1 in 2009Q2**, which was just a one-off impact and should not alter the general trend. It was thus manually adjusted upward to correct the weakness of the statistical projections when there were ad hoc historical figures.
		8. With the **actual** April data from HKTB, and the **monthly seasonality shift in Q2**, we expect that 2011 Q2 should be higher. Year-to-date (YTD) arrivals grew by around **19%** compared to the previous year, and we expect the **momentum** to **continue** in **summer**. Given that the current year is higher than the original forecasts and taking other external factors into consideration, we extend the trend for the rest of the 4 years.

Table 5.6 Comments for China (Continued)

Projections		Comments
		9. Q2 includes the **May holiday** and the forecast should be **upwardly adjusted**. As more individual travellers come to Hong Kong, the **demand** should be **more evenly distributed** throughout the year.
	R2	1. I believe that the forecasts are still valid. There does seem to be improvements over the last two quarters of 2011. However, while the market in China remains strong, it is susceptible to change without much notice, especially with China's inflation problems in their current economic climate. Free independent travelers (FIT) seems to strengthen while group travel remains at a similar levels to the previous year.
		2. As explained by an expert in the previous round, the second quarter arrivals should be adjusted upward.
Pessimistic	R1	1. The growth rate of the **Chinese economy** will **slow down** over the next 5 years, and the **availability** of **new destinations** will **compete** away visitor arrivals from China to Hong Kong.
		2. With the **tax refund policy** in **Hainan** and **more affordable international destinations** for Chinese people, the growth of the Hong Kong market will **slow down.**
	R2	1. There do not seem to be any obvious reasons for the growth of this market to slow down too sharply in the near future (Q2 and Q3 of 2011). Meanwhile, it is believed the **future growth** should **slow down** a bit, and **seasonality** should **not be so obvious** (Q2 arrivals appear relatively lower compared to other seasons).
		2. I think the majority of the experts might have **overestimated** the growth for China. However, as keen competition increases, this market will grow at a **slower** pace.

5.4.3 Arrivals from Japan

As predicted by most experts, inbound demand from Japan was greatly affected as a result of the Japan earthquake (see Table 5.7). It is also likely that Japan's gloomy economic prospects will drag down the growth in arrivals. One expert also mentioned that the aging population in Japan could be another cause for the sluggish growth of tourism demand. This would lead to a decline in arrivals from Japan by 1.21 per cent in 2011.

One expert stated that the group forecasts were too low as “although the earthquake will affect the country’s income level in the short run, Japan will have new opportunities to rebuild its economy and economic growth will increase as a result.” It is anticipated that economic growth in Japan will recover from 2012 onwards and thus gradually boost visitor arrivals from Japan to Hong Kong. In 2015, the number of Japanese visitors to Hong Kong is expected to reach 1.55 million, indicating an increase of 3.55 per cent per annum.

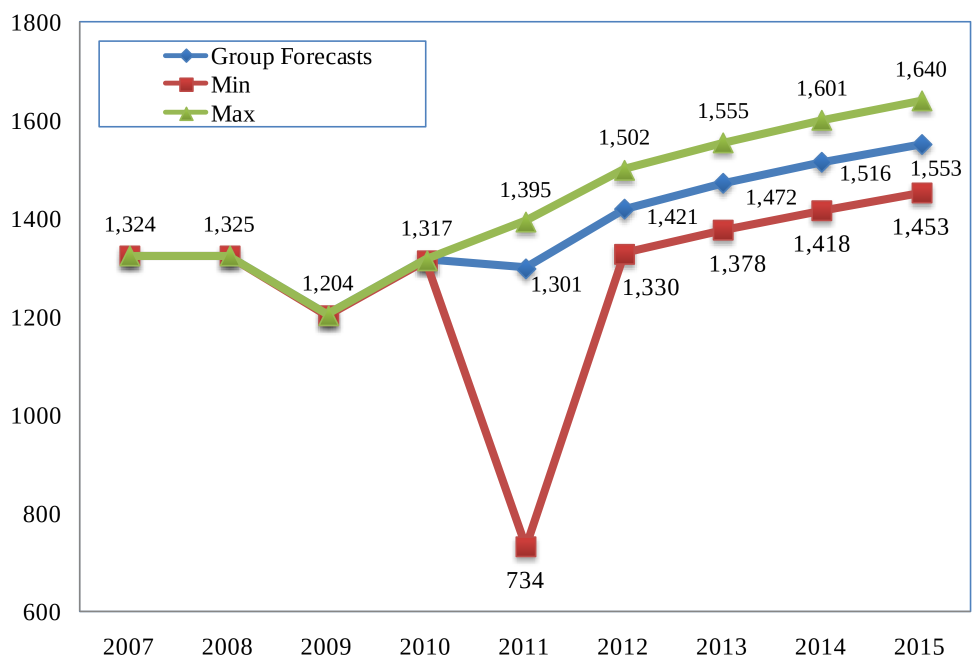


Figure 5.5 Annual forecasts of visitor arrivals from Japan (‘000), 2007−2015

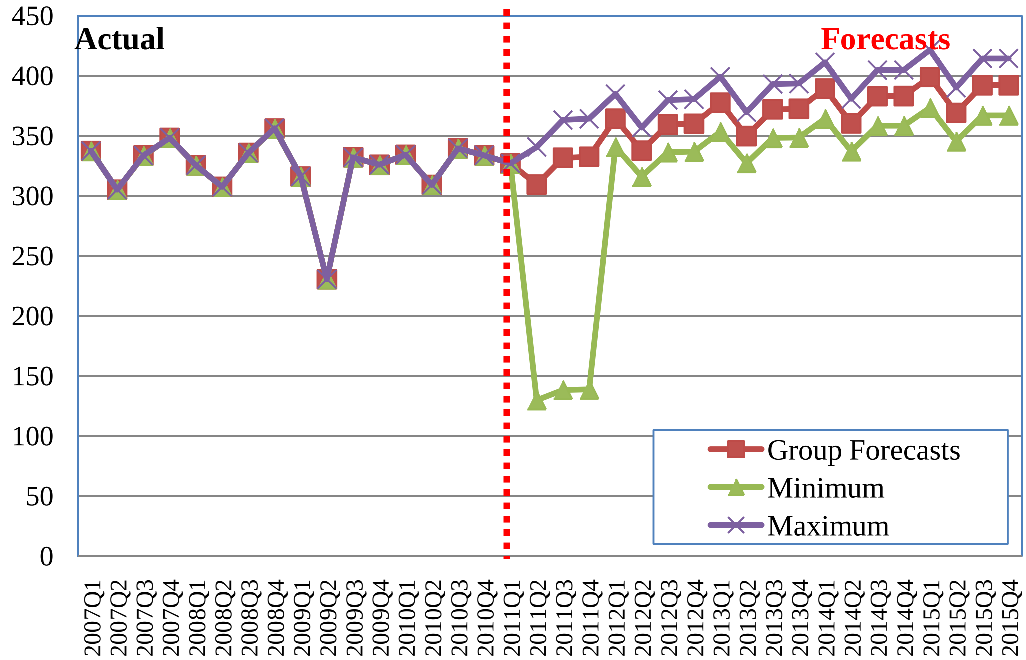


Figure 5.6 Quarterly forecasts of visitor arrivals from Japan (‘000), 2007Q1−2015Q4

Table 5.7 Comments for Japan

Projections		Comments
Optimistic	R1	1. The Japanese market will experience a **slide** in the near future but **growth** will **occur** over the survey period.
	R2	1. The group forecasts were **too low**. Although the earthquake will affect the country’s income level in the short run, Japan will have **new opportunities** to **rebuild its economy** and **economic growth** will **increase** as a result.
Pessimistic	R1	1. The Japan **earthquake** on 11/3 dealt **a heavy blow** to the Japanese market. The number of Japanese outbound tourists had already been **sluggish** over the **past decades**, and the sentiment might be **further dampened** during the reconstruction period in the coming years. The **aging Japan population** is another factor **slowing down** the growth in arrivals.
		2. The Japanese **economy** will **suffer** as a result of the **earthquake** and the **income level** of the residents will be **affected**.
		3. Tourist numbers may **drop dramatically** due to the **economic difficulties** in Japan caused by the **major earthquake** of March 2011.
		4. [It is because of] the **weak economic performance.**
		5. The recent **earthquake** may **disrupt** Japanese tourists’ **outbound travel plans**.
		6. I am afraid that Japan **cannot maintain** the **high speed development** of its **economy**.
		7. **Slow recovery** after the **earthquake** will **dampen** demand for overseas travel in general. Moreover, the Japanese have “seen it all” in Hong Kong.

5.4.4 Arrivals from Taiwan

One expert believed that Hong Kong is still attractive to Taiwanese visitors for shopping and sightseeing. Another expert stated that the Japanese earthquake would have a positive impact on the demand for Hong Kong tourism from Taiwanese tourists over the year 2012. Seasonality was considered by one expert, who stated that the forecasts for the first quarter, which includes the Chinese New Year and winter sales, should be upwardly adjusted (see Table 5.8).

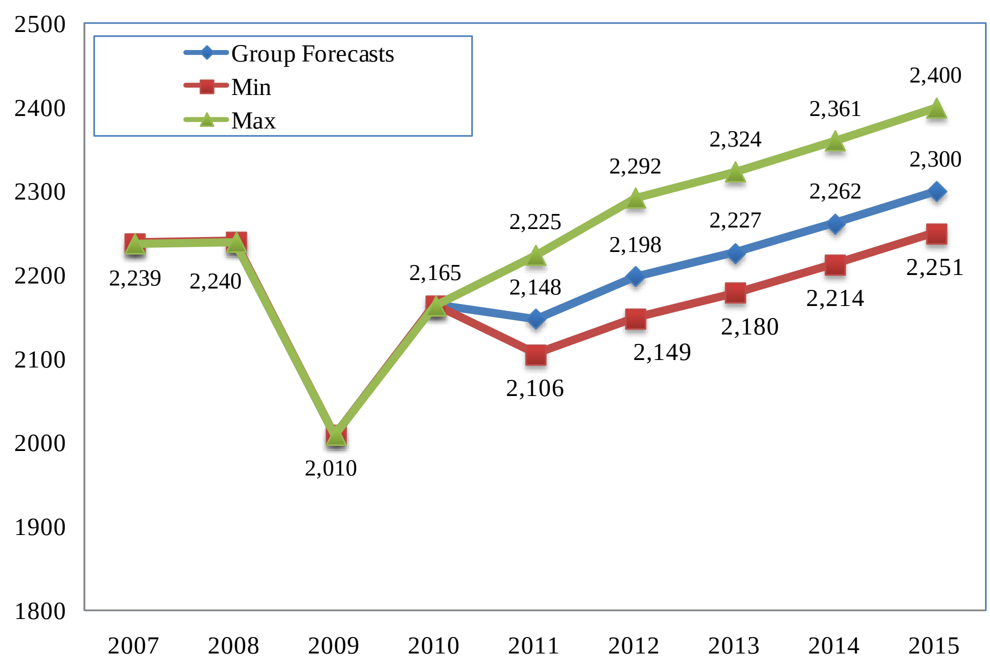


Figure 5.7 Annual forecasts of visitor arrivals from Taiwan (‘000), 2007−2015

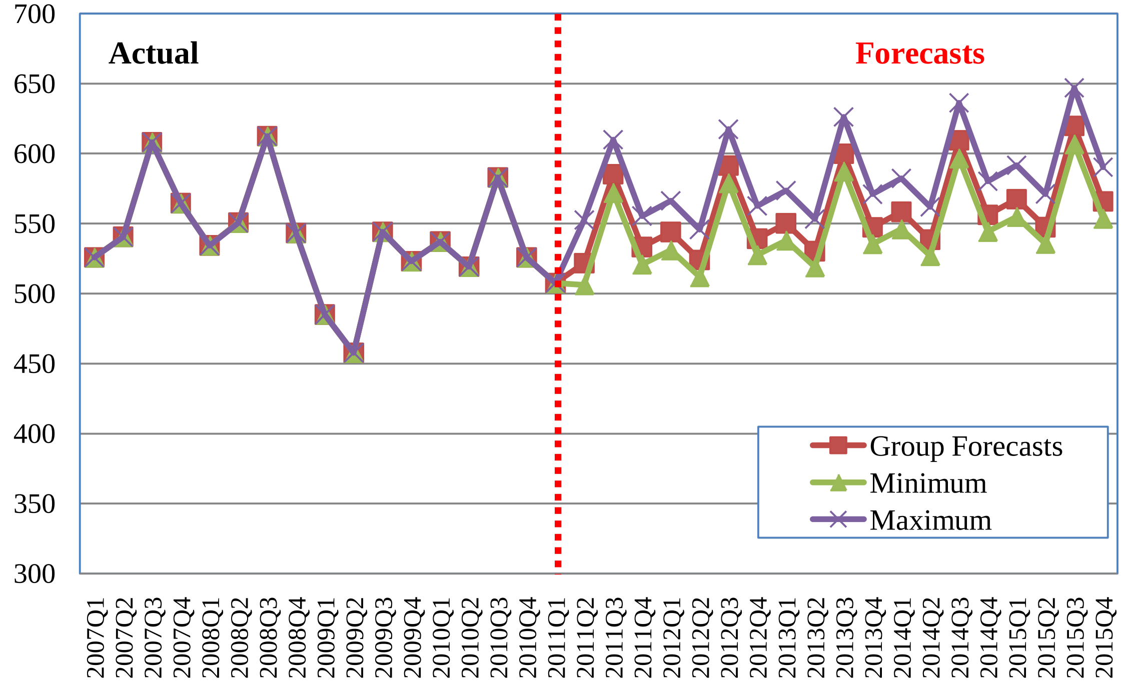


Figure 5.8 Quarterly forecasts of visitor arrivals from Taiwan (‘000), 2007Q1–2015Q4

Three experts stated that the performance of the Taiwanese market will be driven by the relationship between Taiwan and the Mainland. For example, the direct flights to the Mainland, the completion of the high-speed railway, and the increasing socioeconomic interactions between Taiwan and the Mainland will result in Taiwanese tourists switching away from Hong Kong to the Mainland.

In the second round, one expert stated that “[t]he Taiwan market is stable and the current forecasts should remain valid.” The panel believed that Taiwan, Hong Kong’s second largest source market, would not have a strong recovery from the global economic crisis; it was estimated that visitor arrivals from Taiwan would not return to pre-crisis level until 2014. The total number of visitors from Taiwan is expected to reach 2.3 million in 2015, representing an annual increase of 1.22 per cent between

2010 and 2015.

Table 5.8 Comments for Taiwan

Projections		Comments
Optimistic	R1	1. **Shopping and sightseeing** in Hong Kong are still **attractive** activities for Taiwanese tourists and an **increase** in visitor arrivals over **the next five years** is expected. In addition, with the completion of the **Express Rail Link**, some Taiwanese tourists may **travel through China** via Hong Kong.
		2. Hong Kong may **benefit** from the Japan **earthquake** in 2011 over the next year as it is a popular destination for Taiwanese tourists.
		3. **Q1** includes the **Chinese New Year** and **winter sale season**.
Pessimistic	R1	1. With the launch of **direct flights to China** from Taiwan, total arrivals in Hong Kong should **decrease** due to the dropping of en route stop in the territory.
		2. Taiwan market performance will be driven by the relationship with the Mainland and Mainland policies. With the current trend of **increasing interaction**, some Taiwanese visitors might **go to the Mainland instead of** Hong Kong.
		3. With **direct flights** and **closer cooperation** between China and Taiwan, tourists will **switch away** from Hong Kong to China.
		4. As Taiwan is already a **mature** source market for Hong Kong inbound tourism, the growth rate should be **a little bit lower.**
	R2	1. No adjustment was made to this market. I agree with the majority of the experts that the growth should be slower than what I predicted in Round 1 as there will be more direct travel from Taiwan to China in the near future given the relatively positive political situation between the two places (China and Taiwan).

5.4.5 Arrivals from the UK

Table 5.9 shows that the majority of the experts believed that the London Olympic Games would have a negative effect on visitor arrivals from the UK to Hong Kong. One expert expected that the growth of visitor arrivals from the UK would increase in the long run but not substantially. Other experts also agreed that demand would be relatively stable.

Figure 5.9 shows that Hong Kong tourism suffered a greater loss from the UK market (i.e. visitors from the UK declined significantly for the period 2008−2009) due to the global economic turbulence but that a recovery started from 2010 onwards. The UK is projected to be more robust in its recovery than Australia and the USA, with a growth rate of 3.65 per cent per annum from 2010–2015. The number of tourists from the UK will reach 0.62 million in 2015, taking inbound demand back to its pre-crisis level.

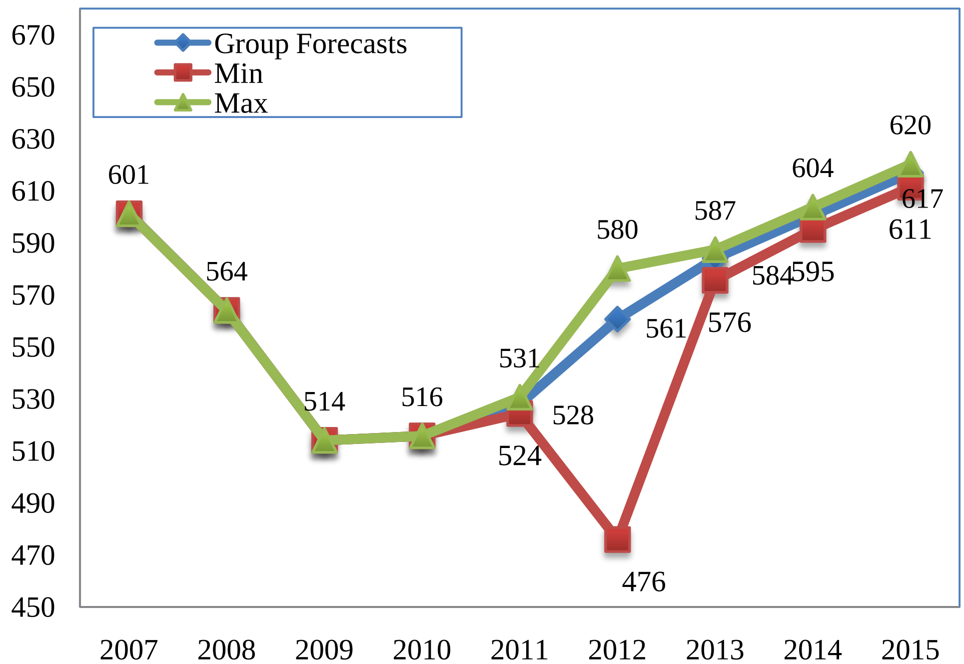


Figure 5.9 Annual forecasts of visitor arrivals from the UK ('000), 2007–2015

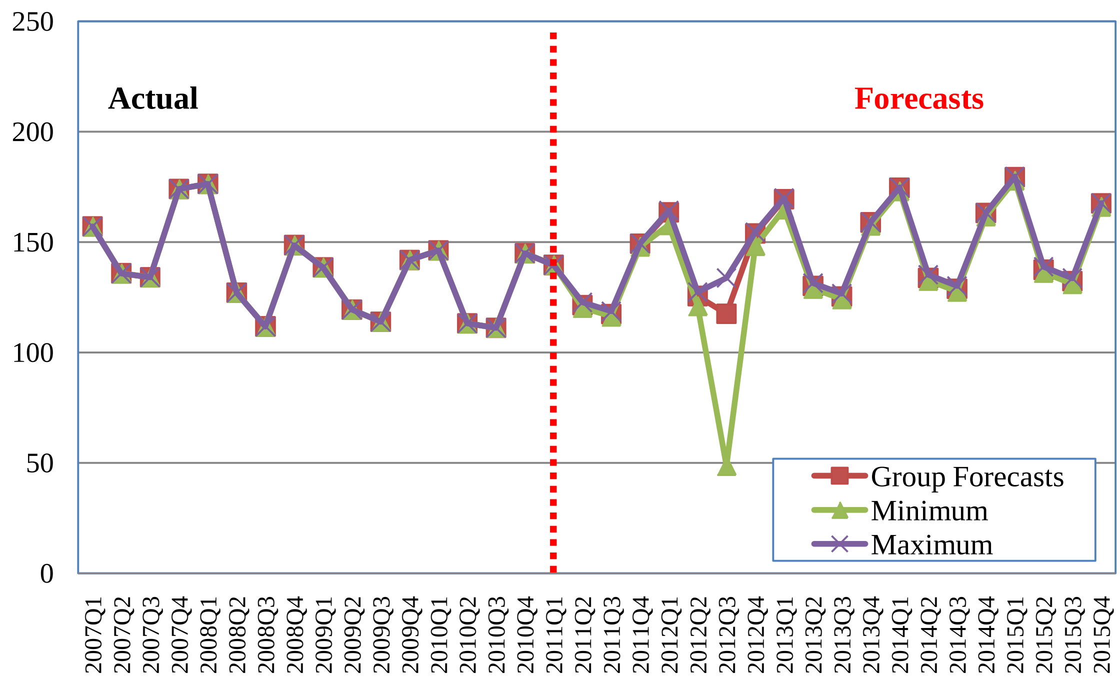


Figure 5.10 Quarterly forecasts of visitor arrivals from the UK ('000), 2007Q1−2015Q4

Table 5.9 Comments for the UK

Projections		Comments
Optimistic	R1	1. Overall, visitor arrivals will **increase** but **not substantially**. Some of the new attractions may account for this, but **Mainland China** will **attract more British** tourists and they will **travel** there **directly**.
		2. I think the demand for Hong Kong tourism from the UK will be **more stable** in the future.
Pessimistic	R1	1. Based on previous experience, hosting the Olympics creates **"stay-home"** effect and **affects outbound** tourism in the year. Thus, arrivals from the UK in **2012** was **adjusted downward.**
		2. The **Departure Tax** introduced in the UK will **hinder** travel to Hong Kong.
		3. The **majority** of British people will **stay** in their country to experience the **Olympic Games** [in 2012].
		4. Overall, the **London Olympics** may have a **negative** effect on British tourists' **outbound travel to Hong Kong**.
	R2	1. The 2012 Q2 and Q3 were **adjusted downwardly** by 2% due to the **"stay home"** effect of the London Olympic Games in 2012.

5.4.6 Arrivals from the USA

Overall, the experts held a pessimistic view on the future growth of US tourist numbers because of the USA's slow economic recovery from the global financial/economic downturn and the weakening US dollar (see Table 5.10). Arrivals to Mainland China from the USA are expected to continue, which might compete away US tourists from visiting Hong Kong. In contrast, however, one expert believed that visitor arrivals from the USA were likely to grow at a moderate pace.

Like the UK market, visitor arrivals from the USA suffered a significant decline from 2008 to 2009 due to the global economic downturn. The recovery has picked up since 2010, and the number of US visitors is expected to reach its pre-crisis level in 2013. The US market is predicted to increase by 2.38 per cent per year over the period 2010−2015.

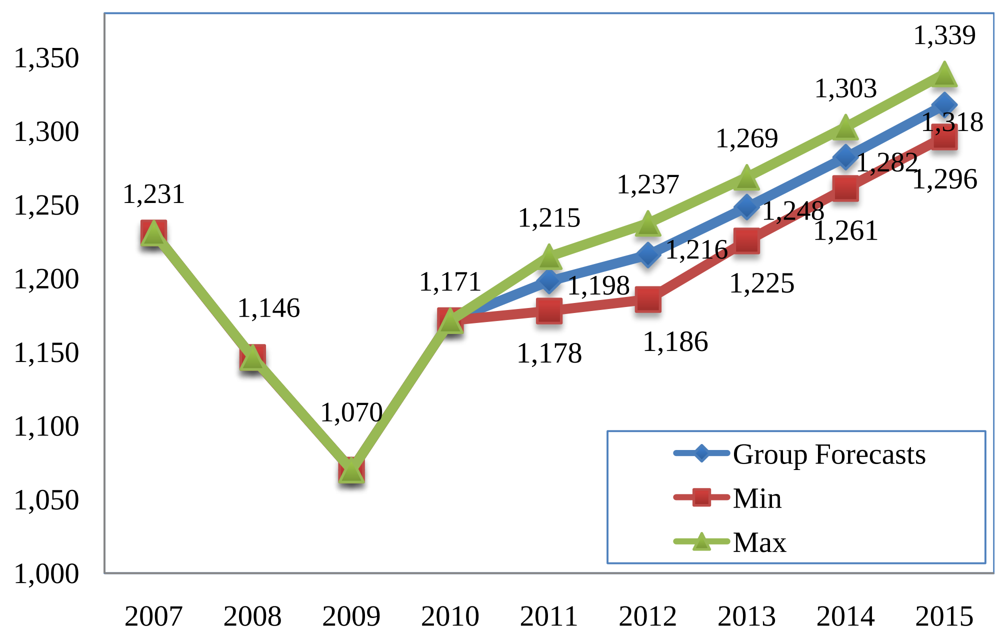


Figure 5.11 Annual forecasts of visitor arrivals from the USA ('000), 2007−2015

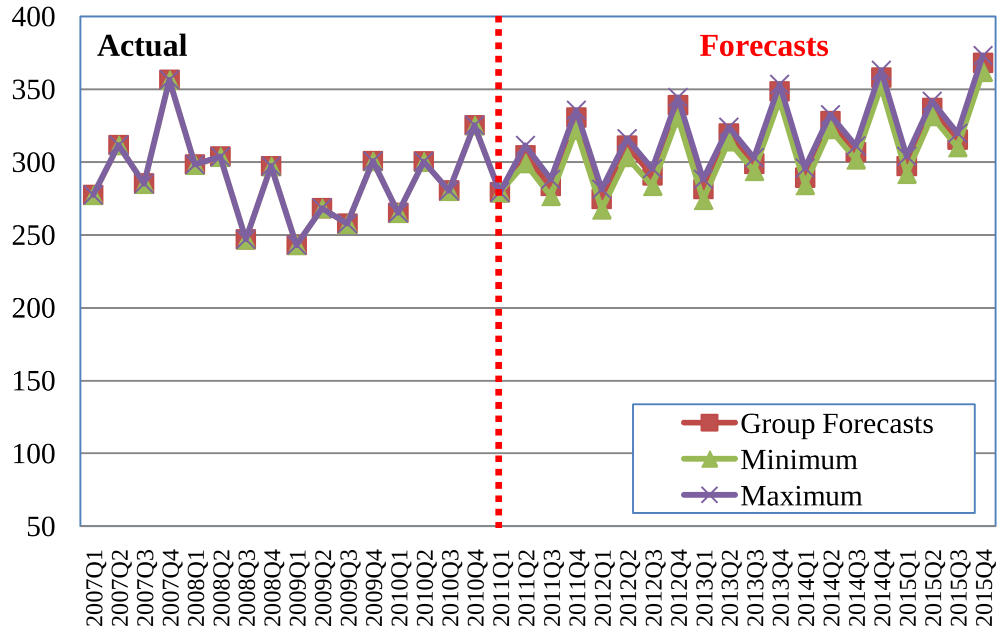


Figure 5.12 Quarterly forecasts of visitor arrivals from the USA ('000), 2007Q1−2015Q4

Table 5.10 Comments for the USA

Projections		**Comments**
Optimistic	R1	1. [It is] because of America's **economic recovery** in the recent future.
		2. The US economy appears to be **proceeding at a moderate pace**, although somewhat more slowly than the Federal Committee had expected. Accordingly, arrivals from the USA are likely to **maintain a moderate pace** of growth.
Pessimistic	R1	1. It seems the US **economic recovery** from the current financial crisis has been **slower than expected**.
		2. The US economy has still **not fully recovered**, and the **weakening US dollar** will also **affect** travel from the USA to Asia.
		3. [It is because of] the **economic downturn**.
		4. We have seen **very little improvement** in the **US economy**. The financial crisis has had a very deep rooted effect, and I do not foresee a recovery in the next 3 years. For 2012, I forecast **a drop** in the tourist market due to it being an **election year**. The business market, however, will improve as orders have been shelved for the last 2 years. Overall, I will see a 2% increase in arrivals for 2012, another 3% for 2013, and a 5% increase for the next 3 years.
		5. While it is expected that there will be **an increase** in the number of US tourists, the **diversion** of these tourists to Mainland China will continue, with demand there growing strongly over the next decade.
	R2	1. [The] overall future trend was **adjusted downward** by 2% due to the **sluggish economic** situation in the USA, which does not suggest recovery in the near future. Hong Kong will lose a large portion of its markets to Mainland destinations as the sky is **opened** for **more direct flights** from major cities in the USA to more destinations in China.

5.5 Evaluation of Forecasting Performance

The APE, MAPE, RMSPE, and Theil's U statistics were used to evaluate forecast accuracy based on a comparison of actual visitor arrivals over the period 2011Q2−2012Q2 and the corresponding forecasts. For the APE, MAPE, and RMSPE measures, a lower value indicates a more accurate forecast. For the U statistic, a value less than one indicates that the statistical/judgmental forecasts are better than the Naive forecasts. Error reductions between each of the two forecasts − statistical forecasts and judgmental forecasts− are presented with accuracy measures in Tables 5.11−5.13. A negative percentage change in either MAPE or RMSPE indicates improvement in accuracy. A detailed analysis was conducted to test the proposed hypotheses according to the results tabulated in Tables 5.11−5.13.

The accuracy measures reported in Table 5.11 are overall averages. A breakdown of these measures is also included for each of the source markets (Table 5.12) and for each quarter (Table 5.13) over different forecasting horizons to give an average accuracy for each round. The findings and implications are discussed in the remainder of this section. Section 5.5.1 reports the accuracy results of the statistical forecasts evaluated by APE, MAPE, U statistic, R^2, and adjusted R^2. Section 5.5.2 presents basic distributional properties of forecast errors. Section 5.5.3 examines the six main hypotheses by evaluating accuracy using different error measures and conducting statistical tests (either parametric or nonparametric tests depending on the results of two tests: normality test and the homogeneity of variance test) to examine the significance of the accuracy difference. Evidence from the existing literature is used to supplement the data analysis and draw implications where appropriate.

Table 5.11 Overall forecasting performance 2011Q2−2012Q2

			MAPE (%)	RMSPE (%)	U	MAPE (%)	RMSPE (%)	U	MAPE (%)	RMSPE (%)	U
Measures	**Group**	**Round**	**Overall**			**Short-haul**			**Long-haul**		
SF			8.59	13.04	1.03	13.92	18.02	1.64	3.25	3.93	0.29
Mean	**All**	GF1	7.54	10.06	0.79	11.14	13.43	1.20	3.94	4.69	0.35
		GF2	6.47	8.56	0.67	9.33	11.40	1.02	3.61	4.05	0.30
	Industry	GF1	6.73	9.08	0.72	10.03	12.22	1.10	3.43	3.93	0.29
		GF2	6.22	8.23	0.64	8.99	10.93	0.97	3.45	4.00	0.30
	Academic	GF1	8.11	10.73	0.84	11.85	14.22	1.27	4.37	5.29	0.39
		GF2	6.74	8.80	0.69	9.78	11.74	1.06	3.70	4.12	0.31
Median	**All**	GF1	7.36	10.14	0.79	11.07	13.74	1.23	3.64	4.12	0.31
		GF2	7.03	9.53	0.74	10.42	12.81	1.14	3.63	4.17	0.31
Self-rated expertise weighted mean	**All**	GF1	7.81	10.42	0.81	11.56	13.90	1.24	4.07	4.91	0.36
		GF2	6.71	8.89	0.69	9.73	11.88	1.06	3.68	4.10	0.31
	Industry	GF1	6.61	8.90	0.70	9.81	11.97	1.08	3.41	3.88	0.29
		GF2	6.25	8.22	0.64	8.97	10.89	0.97	3.53	4.09	0.31
	Academic	GF1	8.68	11.45	0.89	12.59	15.17	1.35	4.78	5.67	0.42
		GF2	6.92	9.21	0.72	10.09	12.34	1.10	3.75	4.15	0.31

Table 5.11 Overall forecasting performance 2011Q2−2012Q2 (Continued)

Measures	Group	Round	Overall MAPE (%)	Overall RMSPE (%)	Overall U	Short-haul MAPE (%)	Short-haul RMSPE (%)	Short-haul U	Long-haul MAPE (%)	Long-haul RMSPE (%)	Long-haul U
% Reduction											
Mean	**All**	GF1-SF	-12.2	-22.9	-23.6	-20.0	-25.5	-26.5	21.2	19.5	18.1
		GF2-SF	-24.6	-34.4	-34.9	-33.0	-36.7	-37.6	11.2	3.2	3.5
		GF2-GF1	-14.1	-15.0	-14.8	-16.2	-15.1	-15.2	-8.2	-13.6	-12.4
	Industry	GF1-SF	-21.6	-30.4	-30.7	-27.9	-32.2	-32.7	5.7	0.1	0.0
		GF2-SF	-27.6	-36.9	-37.8	-35.5	-39.3	-40.7	6.2	1.8	2.3
		GF2-GF1	-7.6	-9.3	-10.3	-10.4	-10.6	-11.8	0.5	1.7	2.3
	Academic	GF1-SF	-5.6	-17.8	-18.8	-14.9	-21.1	-22.4	34.4	34.5	32.1
		GF2-SF	-21.5	-32.5	-32.9	-29.7	-34.8	-35.5	13.7	4.8	4.9
		GF2-GF1	-16.9	-18.0	-17.3	-17.4	-17.4	-16.8	-15.4	-22.1	-20.6
Median	**All**	GF1-SF	-14.3	-22.2	-23.2	-20.5	-23.8	-24.9	12.1	4.9	4.8
		GF2-SF	-18.2	-26.9	-28.3	-25.2	-28.9	-30.6	11.8	6.1	6.1
		GF2-GF1	-4.5	-6.0	-6.7	-5.9	-6.7	-7.6	-0.2	1.1	1.3
Self-rated expertise weighted mean	**All**	GF1-SF	-9.0	-20.1	-21.1	-17.0	-22.8	-24.2	25.2	24.9	23.0
		GF2-SF	-21.9	-31.8	-32.9	-30.1	-34.0	-35.5	13.4	4.4	4.6
		GF2-GF1	-14.2	-14.7	-14.9	-15.8	-14.5	-14.9	-9.4	-16.4	-15.0
	Industry	GF1-SF	-23.0	-31.8	-31.9	-29.5	-33.6	-33.9	4.8	-1.2	-1.3
		GF2-SF	-27.3	-37.0	-38.0	-35.6	-39.6	-41.1	8.6	4.0	4.4
		GF2-GF1	-5.5	-7.6	-8.9	-8.6	-9.0	-10.8	3.6	5.3	5.8
	Academic	GF1-SF	1.1	-12.2	-13.8	-9.6	-15.8	-17.8	47.0	44.4	41.3
		GF2-SF	-19.4	-29.4	-30.5	-27.6	-31.5	-32.9	15.5	5.6	5.6
		GF2-GF1	-20.3	-19.6	-19.3	-19.9	-18.7	-18.4	-21.5	-26.9	-25.3

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

Table 5.12 Forecasting performance by market 2011Q2−2012Q2

Measures	Round	MAPE(%)	RMSPE(%)	U	MAPE(%)	RMSPE(%)	U	MAPE(%)	RMSPE(%)	U
		Australia			**UK**			**USA**		
SF		2.15	2.74	0.33	5.53	5.83	0.36	2.07	2.21	0.15
Mean	GF1	4.03	4.32	0.51	6.37	6.63	0.40	1.41	1.85	0.11
	GF2	3.38	3.57	0.43	5.24	5.57	0.34	2.22	2.36	0.16
Industry	GF1	3.57	3.90	0.46	4.83	5.20	0.32	1.90	2.03	0.14
	GF2	3.33	3.63	0.44	5.19	5.55	0.34	1.84	2.01	0.13
Academic	GF1	4.32	4.64	0.55	7.35	7.56	0.46	1.43	2.25	0.14
	GF2	3.41	3.61	0.43	5.27	5.58	0.34	2.41	2.59	0.18
Median	GF1	3.33	3.48	0.42	5.53	5.83	0.36	2.07	2.21	0.15
	GF2	3.30	3.65	0.44	5.53	5.83	0.36	2.07	2.21	0.15
Self-rated expertise	GF1	4.26	4.62	0.55	6.61	6.85	0.42	1.33	1.97	0.12
weighted mean	GF2	3.51	3.69	0.44	5.24	5.55	0.34	2.31	2.46	0.17
Industry	GF1	3.66	4.04	0.48	4.58	4.95	0.31	1.98	2.09	0.15
	GF2	3.39	3.77	0.46	5.29	5.62	0.35	1.91	2.06	0.14
Academic	GF1	4.62	5.02	0.59	7.81	8.01	0.48	1.91	2.67	0.17
	GF2	3.56	3.74	0.45	5.22	5.52	0.34	2.48	2.68	0.19
		China			**Japan**			**Taiwan**		
SF		28.18	28.78	2.63	8.71	10.65	0.68	4.89	5.66	0.73
Mean	GF1	19.04	19.98	1.82	8.39	9.74	0.63	6.00	6.87	0.88
	GF2	15.69	16.83	1.53	7.26	8.47	0.55	5.05	5.91	0.76
Industry	GF1	17.67	18.54	1.69	6.86	7.99	0.51	5.57	6.40	0.82
	GF2	14.44	15.63	1.42	7.69	9.14	0.59	4.83	5.56	0.72
Academic	GF1	19.91	20.90	1.90	9.36	10.86	0.70	6.27	7.18	0.91
	GF2	16.32	17.44	1.58	7.81	8.49	0.57	5.22	6.10	0.79
Median	GF1	19.61	20.51	1.87	8.71	10.65	0.68	4.89	5.66	0.73
	GF2	17.65	18.63	1.69	8.71	10.65	0.68	4.89	5.66	0.73
Self-rated expertise	GF1	19.39	20.28	1.85	9.19	10.96	0.71	6.09	6.94	0.88
weighted mean	GF2	16.13	17.23	1.57	7.93	9.53	0.61	5.13	6.01	0.77
Industry	GF1	17.42	18.21	1.66	6.52	7.60	0.49	5.50	6.34	0.81
	GF2	14.16	15.36	1.40	7.87	9.37	0.60	4.87	5.64	0.73
Academic	GF1	20.55	21.52	1.96	10.78	13.20	0.86	6.44	7.31	0.93
	GF2	16.99	18.06	1.64	7.95	9.61	0.61	5.32	6.20	0.80

Table 5.12 Forecasting performance by market 2011Q2−2012Q2 (Continued)

Measures	Round	MAPE (%)	RMSPE (%)	U	MAPE (%)	RMSPE (%)	U	MAPE (%)	RMSPE (%)	U
% Reduction		**Australia**			**UK**			**USA**		
Mean	GF1-SF	87.25	57.98	56.27	15.27	13.82	13.09	-31.84	-16.33	-23.92
	GF2-SF	57.07	30.64	31.59	-5.09	-4.45	-4.26	7.11	6.84	8.63
	GF2-GF1	-16.11	-17.31	-15.79	-17.67	-16.05	-15.34	57.15	27.69	42.77
Industry	GF1-SF	65.71	42.61	40.91	-12.53	-10.74	-10.01	-8.06	-8.20	-7.13
	GF2-SF	54.64	32.52	34.54	-6.08	-4.72	-4.21	-11.19	-8.82	-10.71
	GF2-GF1	-6.68	-7.08	-4.52	7.37	6.74	6.45	-3.40	-0.68	-3.86
Academic	GF1-SF	100.95	69.54	67.48	32.96	29.85	28.06	-30.81	1.95	-8.18
	GF2-SF	58.29	32.02	32.32	-4.59	-4.27	-4.25	16.26	17.10	20.40
	GF2-GF1	-21.23	-22.13	-20.99	-28.25	-26.27	-25.23	68.03	14.86	31.12
Median	GF1-SF	54.67	27.29	27.39	0.00	0.00	0.00	0.00	0.00	0.00
	GF2-SF	53.49	33.28	34.37	0.00	0.00	0.00	0.00	0.00	0.00
	GF2-GF1	-0.76	4.70	5.48	0.00	0.00	0.00	0.00	0.00	0.00
Self-rated expertise weighted mean	GF1-SF	97.94	68.85	66.62	19.58	17.66	16.66	-35.57	-10.85	-19.82
	GF2-SF	63.01	34.93	35.62	-5.21	-4.70	-4.59	11.31	11.30	13.87
	GF2-GF1	-17.64	-20.08	-18.60	-20.73	-19.00	-18.22	72.76	24.84	42.02
Industry	GF1-SF	70.09	47.71	45.77	-17.17	-15.07	-14.21	-4.45	-5.31	-3.37
	GF2-SF	57.61	37.67	39.55	-4.35	-3.44	-3.06	-7.99	-6.56	-7.89
	GF2-GF1	-7.33	-6.80	-4.27	15.48	13.69	12.99	-3.71	-1.32	-4.68
Academic	GF1-SF	114.44	83.51	80.71	41.35	37.56	35.27	-7.82	20.72	10.25
	GF2-SF	65.38	36.67	36.69	-5.59	-5.24	-5.26	19.75	21.24	25.19
	GF2-GF1	-22.88	-25.53	-24.36	-33.21	-31.12	-29.96	29.91	0.43	13.55

Table 5.12 Forecasting performance by market 2011Q2‒2012Q2 (Continued)

Measures	Round	MAPE (%)	RMSPE (%)	U	MAPE (%)	RMSPE (%)	U	MAPE (%)	RMSPE (%)	U
% Reduction:		**China**			**Japan**			**Taiwan**		
Mean	GF1-SF	-32.42	-30.59	-30.84	-3.69	-8.60	-7.58	22.71	21.50	19.64
	GF2-SF	-44.30	-41.53	-41.89	-16.60	-20.53	-18.73	3.21	4.47	4.10
	GF2-GF1	-17.58	-15.76	-15.98	-13.40	-13.05	-12.06	-15.89	-14.01	-12.99
Industry	GF1-SF	-37.28	-35.60	-35.73	-21.23	-25.03	-24.25	13.87	13.16	11.71
	GF2-SF	-48.75	-45.71	-45.99	-11.76	-14.22	-13.79	-1.11	-1.65	-1.26
	GF2-GF1	-18.28	-15.70	-15.96	12.02	14.42	13.80	-13.16	-13.09	-11.61
Academic	GF1-SF	-29.32	-27.38	-27.70	7.47	1.91	3.09	28.33	26.91	24.81
	GF2-SF	-42.08	-39.40	-39.81	-10.30	-20.33	-15.99	6.77	7.85	7.14
	GF2-GF1	-18.05	-16.56	-16.75	-16.54	-21.82	-18.50	-16.81	-15.02	-14.16
Median	GF1-SF	-30.40	-28.76	-28.92	0.00	0.00	0.00	0.00	0.00	0.00
	GF2-SF	-37.35	-35.28	-35.82	0.00	0.00	0.00	0.00	0.00	0.00
	GF2-GF1	-9.98	-9.15	-9.70	0.00	0.00	0.00	0.00	0.00	0.00
Self-rated expertise weighted mean	GF1-SF	-31.19	-29.53	-29.73	5.55	2.92	3.95	24.56	22.66	20.64
	GF2-SF	-42.76	-40.15	-40.48	-8.98	-10.58	-10.75	4.90	6.32	5.75
	GF2-GF1	-16.81	-15.07	-15.30	-13.77	-13.12	-14.14	-15.78	-13.32	-12.35
Industry	GF1-SF	-38.18	-36.73	-36.73	-25.17	-28.65	-27.49	12.57	12.02	10.67
	GF2-SF	-49.75	-46.64	-46.91	-9.70	-12.05	-11.54	-0.35	-0.36	-0.16
	GF2-GF1	-18.71	-15.67	-16.09	20.68	23.27	21.99	-11.48	-11.05	-9.79
Academic	GF1-SF	-27.05	-25.23	-25.56	23.76	23.88	25.92	31.66	29.26	26.88
	GF2-SF	-39.70	-37.26	-37.63	-8.67	-9.76	-10.11	8.71	9.53	8.66
	GF2-GF1	-17.34	-16.08	-16.22	-26.21	-27.16	-28.61	-17.43	-15.26	-14.37

Table 5.13 Forecasting performance by forecasting horizon 2011Q2−2012Q2

Measures	Group	Round	MAPE(%)	RMSPE(%)	U	MAPE (%)	RMSPE (%)	U	MAPE (%)	RMSPE (%)	U
			***h* = 1**			***h* = 2**			***h* = 3**		
SF			10.64	14.56	1.19	7.01	11.53	0.96	6.53	9.04	0.56
Mean	**All**	GF1	7.44	10.12	0.82	6.34	8.54	0.69	6.21	6.92	0.44
		GF2	6.93	8.46	0.69	4.35	6.77	0.55	5.7	5.9	0.39
	Industry	GF1	6.7	8.74	0.71	5.48	7.79	0.63	5.85	6.4	0.41
		GF2	6.81	8.69	0.69	5.06	6.67	0.53	4.78	5.14	0.34
	Academic	GF1	8.05	11.05	0.89	6.89	9.04	0.72	6.43	7.3	0.47
		GF2	6.99	8.46	0.7	4.69	7	0.57	6.16	6.42	0.42
Median	**All**	GF1	8.72	11.51	0.92	6.01	8.56	0.7	5.41	6.23	0.4
		GF2	8.45	10.7	0.85	5.14	7.3	0.59	5.28	5.99	0.39
Self-rated expertise weighted mean	**All**	GF1	7.94	11.1	0.88	7.09	8.97	0.72	5.9	6.97	0.45
		GF2	7.6	9.61	0.77	4.83	7.02	0.57	5.36	5.64	0.37
	Industry	GF1	6.36	8.29	0.68	5.28	7.73	0.63	5.95	6.45	0.41
		GF2	6.67	8.64	0.68	5.06	6.61	0.53	4.81	5.18	0.35
	Academic	GF1	9.21	12.91	1.02	8.15	9.85	0.79	6	7.45	0.48
		GF2	8.01	10.08	0.81	4.8	7.24	0.59	5.6	5.88	0.38
			***h* = 4**			***h* = 5**					
SF			8.77	12.29	1.4	9.98	16.51	1.48			
Mean	**All**	GF1	7.98	9.71	1.1	9.73	13.73	1.23			
		GF2	7.06	8.35	0.94	8.32	12.00	1.07			
	Industry	GF1	6.98	8.65	0.98	8.66	12.63	1.13			
		GF2	6.68	7.95	0.9	7.77	11.37	1.02			
	Academic	GF1	8.62	10.44	1.18	10.55	14.46	1.29			
		GF2	7.26	8.56	0.97	8.59	12.32	1.10			
Median	**All**	GF1	7.71	9.51	1.08	8.93	13.39	1.20			
		GF2	7.59	9.27	1.05	8.67	12.81	1.14			
Self-rated expertise weighted mean	**All**	GF1	8.12	9.9	1.12	10.02	13.90	1.24			
		GF2	7.23	8.55	0.96	8.50	12.19	1.09			
	Industry	GF1	6.92	8.58	0.97	8.53	12.33	1.10			
		GF2	6.82	8.03	0.91	7.87	11.33	1.01			
	Academic	GF1	9.01	10.76	1.22	11.05	14.87	1.32			
		GF2	7.42	8.78	0.99	8.78	12.57	1.13			

Table 5.13 Forecasting performance by forecasting horizon 2011Q2−2012Q2 (Continued)

Measures	Group	Round	MAPE (%)	RMSPE (%)	U	MAPE (%)	RMSPE (%)	U	MAPE (%)	RMSPE (%)	U
			h = 1			*h* = 2			*h* = 3		
Mean	**All**	GF1-SF	-30.1	-30.5	-31.4	-9.6	-26	-28.7	-5	-23.5	-20.2
		GF2-SF	-34.8	-41.9	-42.2	-37.9	-41.3	-43.1	-12.7	-34.8	-30.6
		GF2-GF1	-6.8	-16.4	-15.7	-31.3	-20.7	-20.1	-8.1	-14.8	-13
	Industry	GF1-SF	-37	-40	-40.2	-21.9	-32.5	-34.2	-10.4	-29.2	-26.9
		GF2-SF	-36	-40.3	-41.9	-27.9	-42.2	-44.4	-26.9	-43.2	-38.2
		GF2-GF1	1.7	-0.6	-2.9	-7.7	-14.4	-15.5	-18.4	-19.7	-15.3
	Academic	GF1-SF	-24.3	-24.1	-25.5	-1.8	-21.6	-25	-1.6	-19.3	-15.3
		GF2-SF	-34.3	-41.9	-41.5	-33	-39.3	-40.8	-5.7	-29	-25.2
		GF2-GF1	-13.2	-23.4	-21.5	-31.8	-22.5	-21.1	-4.2	-12.1	-11.6
Median	**All**	GF1-SF	-18	-21	-22.5	-14.3	-25.8	-27.1	-17.2	-31.1	-28
		GF2-SF	-20.6	-26.5	-28.4	-26.7	-36.7	-38.2	-19.2	-33.7	-30.3
		GF2-GF1	-3.1	-7	-7.6	-14.5	-14.7	-15.2	-2.5	-3.8	-3.2
Self-rated expertise weighted mean	**All**	GF1-SF	-25.4	-23.8	-25.7	1.1	-22.2	-25.2	-9.7	-22.9	-19.7
		GF2-SF	-28.6	-34	-35.5	-31	-39.1	-40.8	-18	-37.6	-33.6
		GF2-GF1	-4.3	-13.4	-13.2	-31.8	-21.7	-20.8	-9.1	-19.1	-17.3
	Industry	GF1-SF	-40.2	-43.1	-43	-24.6	-33	-34.7	-9	-28.6	-26.5
		GF2-SF	-37.3	-40.6	-42.5	-27.8	-42.7	-45	-26.4	-42.8	-37.7
		GF2-GF1	4.9	4.3	0.8	-4.2	-14.5	-15.9	-19.1	-19.8	-15.2
	Academic	GF1-SF	-13.4	-11.3	-14.4	16.3	-14.6	-18.3	-8.2	-17.6	-13.5
		GF2-SF	-24.8	-30.8	-32.1	-31.6	-37.2	-38.6	-14.3	-35	-31.4
		GF2-GF1	-13.1	-21.9	-20.7	-41.2	-26.5	-24.9	-6.6	-21.1	-20.6

Table 5.13 Forecasting performance by forecasting horizon 2011Q2−2012Q2 (Continued)

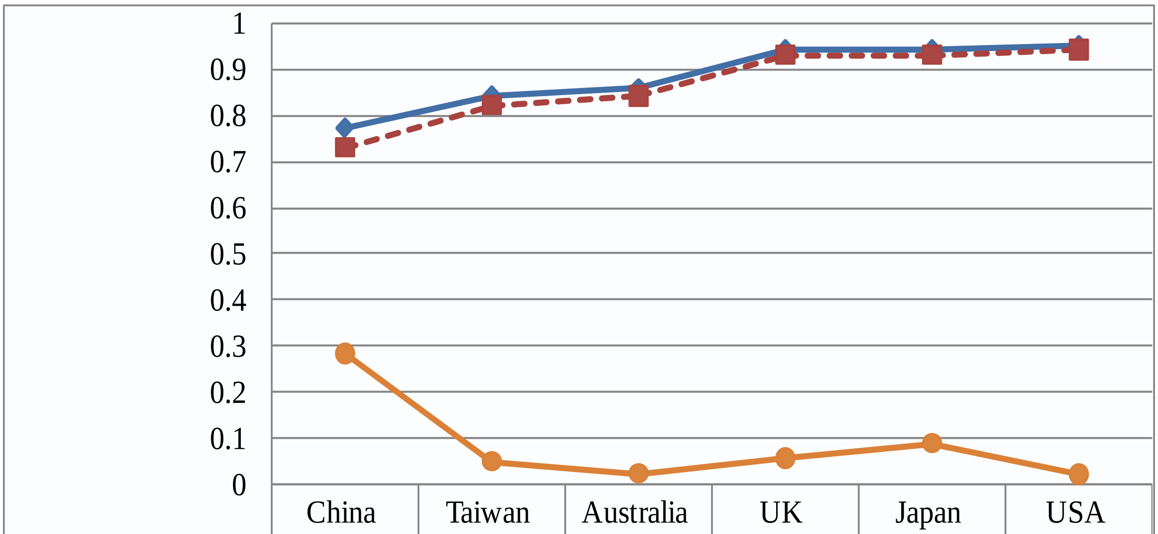
			***h* = 4**			***h* = 5**			
Mean	**All**	GF1-SF	-9	-21	-21.4	-2.5	-16.8	-17.3	
		GF2-SF	-19.5	-32.1	-32.6	-16.6	-27.3	-27.6	
		GF2-GF1	-11.5	-14	-14.2	-14.5	-12.6	-12.4	
	Industry	GF1-SF	-20.5	-29.6	-30.2	-13.2	-23.5	-23.7	
		GF2-SF	-23.9	-35.3	-35.8	-22.1	-31.1	-31.5	
		GF2-GF1	-4.2	-8.2	-8	-10.3	-9.9	-10.2	
	Academic	GF1-SF	-1.7	-15.1	-15.5	5.7	-12.4	-13.1	
		GF2-SF	-17.3	-30.4	-31	-13.9	-25.3	-25.6	
		GF2-GF1	-15.9	-18	-18.4	-18.6	-14.8	-14.4	
Median	**All**	GF1-SF	-12.1	-22.6	-22.9	-10.5	-18.9	-19.2	
		GF2-SF	-13.5	-24.6	-24.9	-13.1	-22.4	-22.8	
		GF2-GF1	-1.6	-2.5	-2.6	-2.9	-4.3	-4.4	
Self-rated expertise weighted mean	**All**	GF1-SF	-7.4	-19.5	-19.9	0.4	-15.8	-16.3	
		GF2-SF	-17.6	-30.5	-31.1	-14.8	-26.2	-26.4	
		GF2-GF1	-10.9	-13.6	-13.9	-15.1	-12.3	-12.1	
	Industry	GF1-SF	-21.1	-30.2	-30.8	-14.5	-25.3	-25.5	
		GF2-SF	-22.3	-34.7	-35.2	-21.1	-31.4	-31.8	
		GF2-GF1	-1.5	-6.4	-6.3	-7.7	-8.1	-8.5	
	Academic	GF1-SF	2.7	-12.5	-12.9	10.8	-9.9	-10.6	
		GF2-SF	-15.5	-28.6	-29.2	-12.0	-23.8	-24.0	
		GF2-GF1	-17.7	-18.3	-18.7	-20.6	-15.5	-15.0	

5.5.1 Accuracy of statistical forecasts

Before exploring the effectiveness of judgmental forecasting on the basis of the statistical forecasting model, it is first necessary to examine the accuracy of the forecasts produced by the econometric model. This study employed the *ex ante* approach which incorporates the model builder's assumptions about the exogenous variables. Thus, in order to evaluate the quality of an econometric model per se, it is necessary to eliminate the effects of the model builder's assumptions, as suggested by Stekler (2007). However, it is extremely difficult to eliminate such impacts in real forecasting practice. Hence, part of the forecast errors probably result from the error associated with projections from the explanatory variables.

The assessment of the likely forecasting ability of econometric models of tourism demand on the basis of common criteria such as goodness of fit, statistical significance of the coefficients, and the like may well be misleading (Witt & Witt, 1992). Conditions such as a high goodness of fit and a large proportion of statistically significant coefficients are insufficient to ensure a high level forecast accuracy. Armstrong (2001b) argued that high R^2 does not ensure accurate forecasts. However, there is certainly some validity in using the fit of a model as a guide to forecast accuracy. For example, a model that cannot explain large historical variations is unlikely to be useful in forecasting. The R^2 value cannot only serve as a measure for the model fit, but also as a judgment of forecast accuracy. The use of R^2 may be useful for evaluating *ex ante* forecasts in this study as the forecasts were made without any knowledge of what happened in the actual situation.

Three models fit the data exceptionally well based on their high R^2 values of 0.95 (USA), 0.94 (UK), and 0.94 (Japan). As shown in Figure 5.13, five out of the six models produced highly accurate arrival forecasts with very small mean and median MAPE values. Among the six source markets, five recorded a mean MAPE of less than 10 per cent and three of them even had a value of less than 5 per cent (see Table 5.14). The largest forecast errors evaluated by mean MAPE were found in the China model, followed by Japan. The U statistics show that all of the models except for the China model outperformed the Naive 1 model.



	China	Taiwan	Australia	UK	Japan	USA
R2	0.77	0.84	0.86	0.94	0.94	0.95
Adjusted R2	0.73	0.82	0.84	0.93	0.93	0.94
Mean MAPE	28.18%	4.89%	2.15%	5.53%	8.71%	2.07%

Figure 5.13 The relationship between R^2 and MAPE by market 2011Q2–2012Q2

Table 5.14 shows that the forecast errors evaluated by APE for an average of one- to five-quarter-ahead forecasts were significantly related to the degree of goodness-of-fit, $r = -0.70$, p (one-tailed) < 0.01: the lower the APE, the higher the R^2. This implies that statistical models with a higher R^2 are likely to produce forecasts that are more accurate.

Table 5.14 Accuracy of statistical forecasts 2011Q2−2012Q2

Test statistic	**Australia**	**China**	**Japan**	**Taiwan**	**UK**	**USA**
R^2	0.86	0.77	0.94	0.84	0.94	0.95
Adjusted R^2	0.84	0.73	0.93	0.82	0.93	0.94
APE (%)						
2011Q2	5.01	28.82	19.25	5.47	3.35	1.93
2011Q3	1.35	26.88	4.42	0.38	7.03	2.00
2011Q4	2.99	20.00	1.92	4.39	7.16	2.75
2012Q1	0.16	26.87	6.49	9.30	6.90	2.92
2012Q2	1.25	38.30	11.46	4.91	3.19	0.76
MAPE (%)	2.15	28.18	8.71	4.89	5.53	2.07
Median APE (%)	1.35	26.88	6.49	4.91	6.90	2.00
RMPSE (%)	2.74	28.78	10.65	5.66	5.83	2.21
Theil's U	0.33	2.63	0.68	0.73	0.36	0.15
Pearson Correlation (APE, R^2)						-0.70^{**}

Note: Correlation is significant at the 1% level (1-tailed).

5.5.2 Basic distributional properties of forecast errors

As presented in Section 3.7 of Chapter 3, standard statistical procedures were applied for forecast evaluation along the dimensions of unbiasedness, efficiency, and accuracy. Prior to the hypothesis testing, it is necessary to present an overall statistical summary of the forecast errors to provide basic quantitative information on the forecasting performance.

Table 5.15 offers a simple view of the distribution of forecast errors (measured by PE) for the arrival series studied. From Table 5.15, it can be seen that there is statistically significant evidence that the percentage errors are not normally distributed as the *p* values of S-W statistic are greater than 0.05. To test mean- and median-unbiasedness, the student one-sample *t* test and the one-sample Wilcoxon signed-rank test were used respectively. These two test results indicate that the percentage errors were unbiased as the corresponding *p* values were higher than 0.05.

Table 5.15 Central tendencies and other distributional properties

PE	SF	$GF1_{mean}$	$GF2_{mean}$	$GF1_{median}$	$GF2_{median}$
Mean	2.22%	-0.32%	0.45%	0.46%	0.27%
Mean (test if 0) [(1)]					
t statistic	0.93	-0.17	0.28	0.25	0.15
p value (2-tailed)	0.36	0.87	0.78	0.81	0.88
5% Trimmed Mean	1.44%	-1.00%	-0.09%	-0.03%	-0.17%
Median	-1.30%	-3.14%	-1.85%	-2.59%	-1.92%
Median (test if 0) [(2)]					
p value (2-tailed)	0.688	0.237	0.750	0.371	0.504
Variance	0.02	0.01	0.01	0.01	0.01
Std. Deviation	0.13	0.10	0.09	0.10	0.10
Minimum	-19.25%	-15.28%	-13.01%	-19.25%	-19.25%
Maximum	38.30%	29.85%	26.46%	29.97%	28.43%
Range	57.56%	45.13%	39.48%	49.23%	47.69%
Interquartile Range	8.75%	9.33%	9.45%	8.38%	8.86%
Skewness	1.37	1.33	1.14	1.13	1.00
Kurtosis	1.51	1.69	1.60	1.63	1.68
Normality test					
S-W statistic	0.84	0.88	0.92	0.89	0.92
p value (2-tailed)	0.00	0.00	0.03	0.01	0.03

Note: (1) Test results are based on the one-sample student *t* test; (2) Test results are based on the one-sample Wilcoxon signed-rank test; (3) SF: statistical forecasts (baseline forecasts), $GF1_{mean}$: consensus forecasts by mean in R1; $GF2_{mean}$: consensus forecasts by mean in R2; $GF1_{median}$: consensus forecasts by median in R1; $GF2_{median}$: consensus forecasts by median in R2.

It is worth noting that the above two bias tests should be interpreted in a cautious manner due to the concern that these cross-market forecast errors may not be drawn from the same underlying population. Specifically, where bias is not found, it could be that bias nevertheless exists but only in some individual markets. On the other hand, where bias is found, this could be misleading as it may only reflect a strong bias from a particular market.

5.5.3 Results of hypothesis testing

To test the research hypotheses presented in Section 3.6 of Chapter 3, this study evaluated forecasting evaluation from three dimensions, namely accuracy, bias, and efficiency. Section 3.7 of Chapter 3 provides methodological details on how to evaluate forecasting performance and conduct statistical tests where appropriate. Specifically, this study employed a group of error measures to examine forecast accuracy and used associated statistical tests to examine whether the accuracy difference was statistically significant. The investigation of potential forecast bias and efficiency was carried out by conducting regression analysis. A traditional comparison to a Naive forecast was also made.

(1) Hypotheses on the accuracy of judgmentally adjusted forecasts

H1a: Judgmental forecast adjustments based on statistical forecasts improve forecast accuracy.

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 5.11, the judgmentally adjusted forecasts were more accurate than the statistical forecasts alone (i.e. baseline forecasts): the mean MAPE decreased from 8.59 to 7.54 per cent in the initial round (R1) and to 6.47 per cent in the subsequent round (R2). The percentage reductions of MAPE ranged from 9.0 to 24.6 per cent, 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 a significant improvement in the forecast accuracy. It was also found that there was no big difference when using MAPE and RMSPE to evaluate the forecast accuracy. Table 5.11 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).

Wilcoxon signed-rank tests were applied to examine if any significant difference existed between forecasts before and after adjustment. The results in Table 5.16 show that the statistical forecasts did not significantly outperform the statistical group forecasts (Z = -0.463, p = 0.328; T =17, r = -0.06). However, compared to the initial statistical forecasts (Round 1 forecasts), the forecast accuracy measured by APE was found to produce more accurate Round 2 forecasts: the p values were 0.099 and 0.004 respectively for Test 2 and Test 3, which were both statistically significant at the 10% level.

Table 5.16 Wilcoxon signed rank test results evaluated by APE

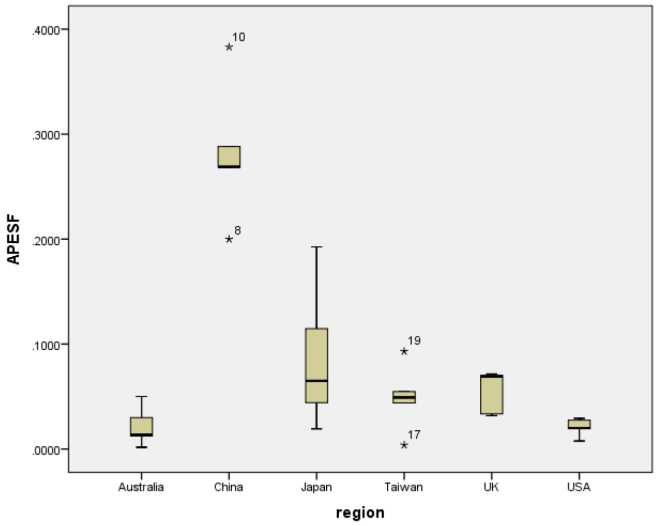
H_0: test if 0 H_1: test if <0	Test 1 ($APE_{GF1}-APE_{SF}$)	Test 2 ($APE_{GF2}-APE_{SF}$)	Test 3 ($APE_{GF2}-APE_{GF1}$)
Positive ranks (T)	17	12	7
Z	-0.463	-1.306	-2.643
Exact p. (1-tailed)	**0.328**	**0.099**	**0.004**
Effect size (r)	-0.06 (Small effect)	-0.17 (Small effect)	-0.34 (Medium effect)

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 Tables 5.11−5.13). Table 5.12 suggests that the largest accuracy improvement over the statistical forecasts was found in the prediction of visitor arrivals from the Mainland, 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

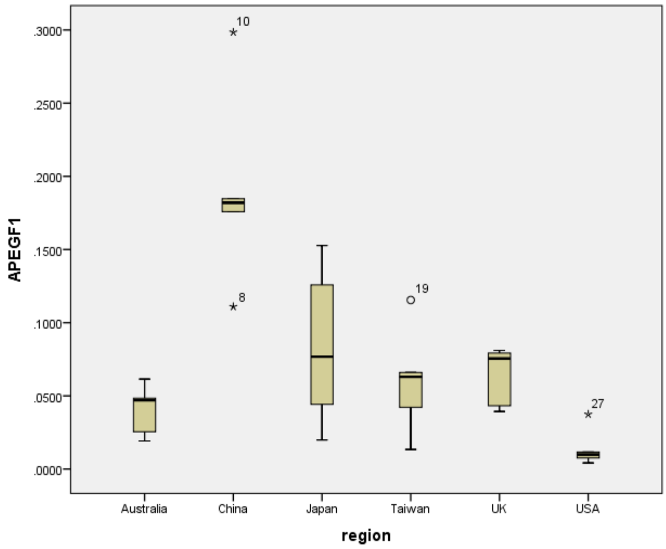
3%). When similar comparisons to those shown in Table 5.12 were made using APE, the results were found to be similar in most cases.

Table 5.17 provides a more detailed analysis of the performance statistics for individual quarters by markets and rounds as assessed by APE. Figure 5.14 shows the distribution of APEs split by market. The APEs of the three sets of forecasts (SF, GF1, and GF2) were calculated for each quarter between 2011Q2 and 2012Q2.

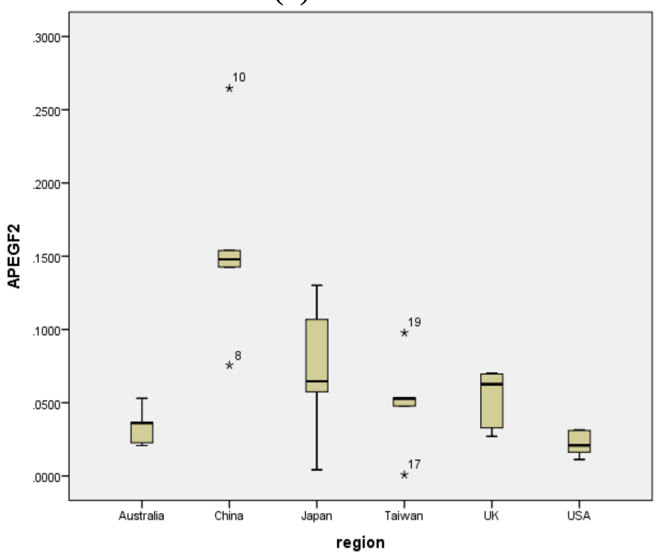
The cumulative frequencies of the negative error differences between the two forecasts are presented in Table 5.17 as percentages. Reductions in APE or an improvement in forecast accuracy as a result of using the forecasting adjustment method (versus statistical forecasting alone) were observed in five 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 the pilot study presented in Chapter 4, this confirmed that forecasts for the long-haul markets were more accurate than those for the short-haul markets.



(a) SF



(b) R1



(c) R2

Figure 5.14 Boxplots of APEs by market

Table 5.17 Forecasting performance evaluated by APE (%) by market

Country/ Region	CV	Quarter	APE_{SF} (a)	APE_{GF1} (b)	APE_{GF2} (c)	$PB_{(b-a<0)}$	$PB_{(c-a<0)}$	$PB_{(c-b<0)}$
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			
Medium	0.41	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			
High	1.14	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			
Medium	0.32	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			
Medium	0.38	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			
Medium	0.33	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			
Low	0.26	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.

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 1.92 to 19.25 per cent. 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 per cent.

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 *H1a*.

H1b: Judgmentally adjusted forecasts are more accurate than Naive forecasts.

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 Naive 1 were considered by calculating the U statistic. The overall performance of the statistical forecasts was as similar to Naive 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 *H1b* that, on average, judgmentally adjusted forecasts are more accurate than Naive 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 5.12 shows that the high value of the U statistic was mainly due to the Mainland market, which had a relatively large value (2.16 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 Naive forecasts. Another possible reason for the larger U statistics is the length of forecasting horizon for calculating the U statistic. Table 5.13 shows clearly that the value of the U statistic was higher in the subsamples with longer horizons; for example, for h = 5, the U statistics were above one for the three sets of forecasts; but for h = 2 and h = 3, the U statistics were far below one.

(2) Hypotheses on the bias and inefficiency of judgmentally adjusted forecasts

H2a: Judgmentally adjusted forecasts are biased.

The literature shows that forecasts produced by models are better than unaided judgment, but on the other hand, the literature also illustrates that the use of judgment can introduce biases (Stekler, 2007). People's predictions will therefore contain at least some component of errors (Armor & Taylor, 2002). It is important to know both the effects and biases that can result from such judgmental predictions.

To test for the bias of the judgmentally adjusted forecasts, a pooled regression model of Equation (3.7) 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 5.19 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 adjustment forecasts for R1 and R2 were unbiased: α was insignificant, indicating that there was, on average, no bias in the forecasts either in the first or second round.

To investigate whether or not the adjusted forecasts made by individual experts were biased, Equation (3.7) 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 statistically indifferent from zero suggesting that the individual experts' adjusted forecasts were unbiased.

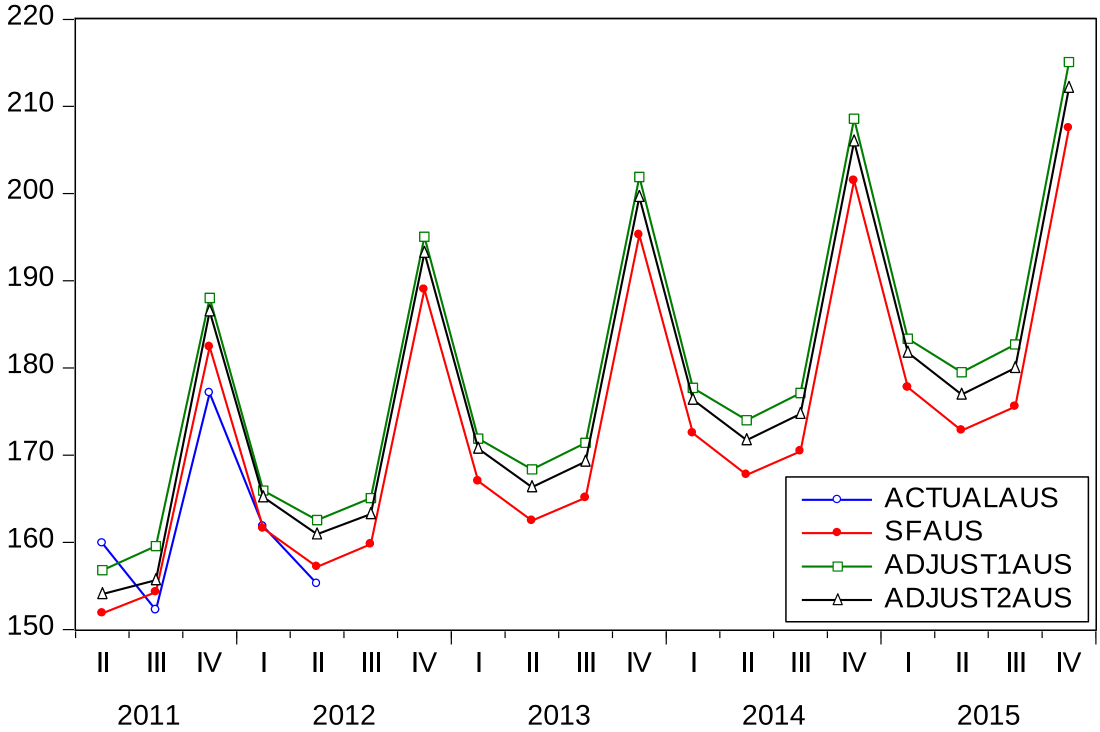
In addition to the regression analysis, an alternative test of forecast bias – the percentage of cases where the forecast (either adjusted or unadjusted) was greater than the actual value was computed and the binomial test was used to determine whether this was significantly different from 0.5 (50%). The binomial tests shown in Table 5.18 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.

Table 5.18 Binomial test results (bias is measured by the number of (F>A) and (F<A))

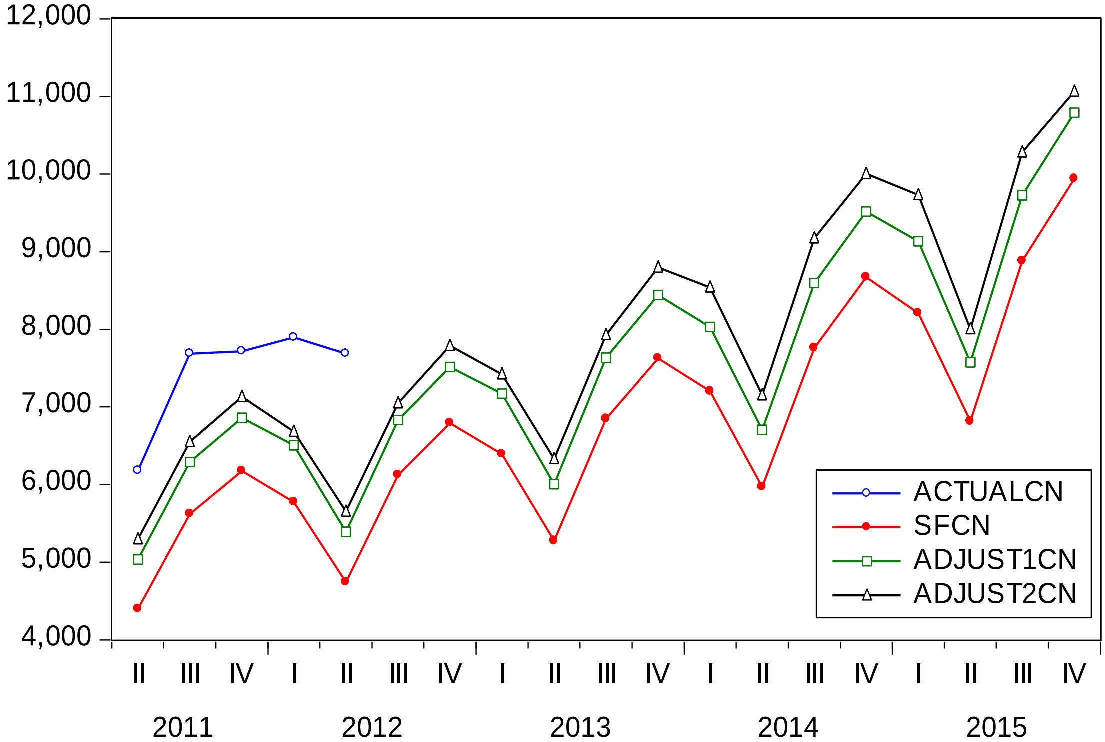
	Category	N	Observed Proportion	*p* (2-tailed)
SF	F<A	14	0.47	0.856
	F>A	16	0.53	
	Total	30	1.00	
GF1	F<A	14	0.47	0.856
	F>A	16	0.53	
	Total	30	1.00	
GF2	F<A	11	0.37	0.200
	F>A	19	0.63	
	Total	30	1.00	

Even though the regression analysis and the binominal test results both suggested that three sets of forecasts − the statistical forecasts and the group forecasts in Rounds 1 and ~~2~~ were unbiased, it should 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). As shown in Appendix I (in Appendix A), there was a mixture of different trends in the six source 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.

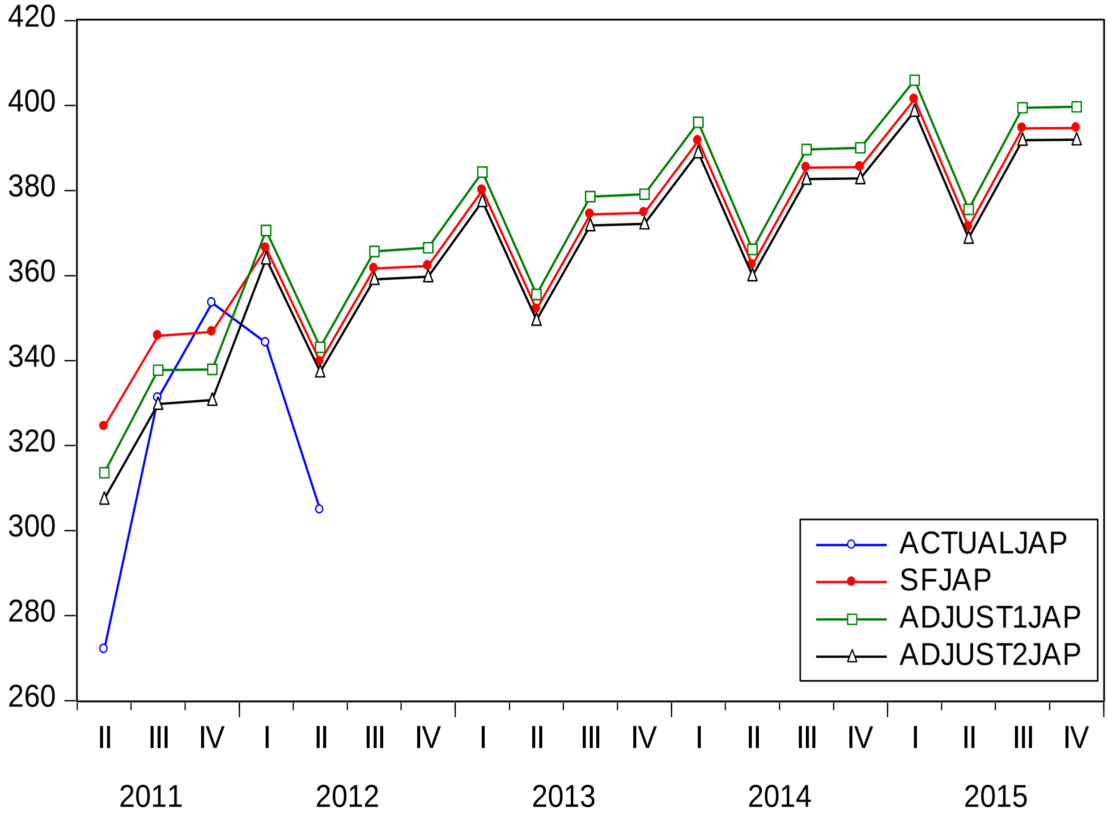
A closer analysis of the individual market results revealed that the majority of the forecasts overestimated future arrivals. Figure 5.15 provides visual evidence of the direction of the bias for individual markets. It can be seen from Figure 5.15 that forecasts from Australia, Taiwan, and the UK were overestimated while forecasts from the Mainland 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.



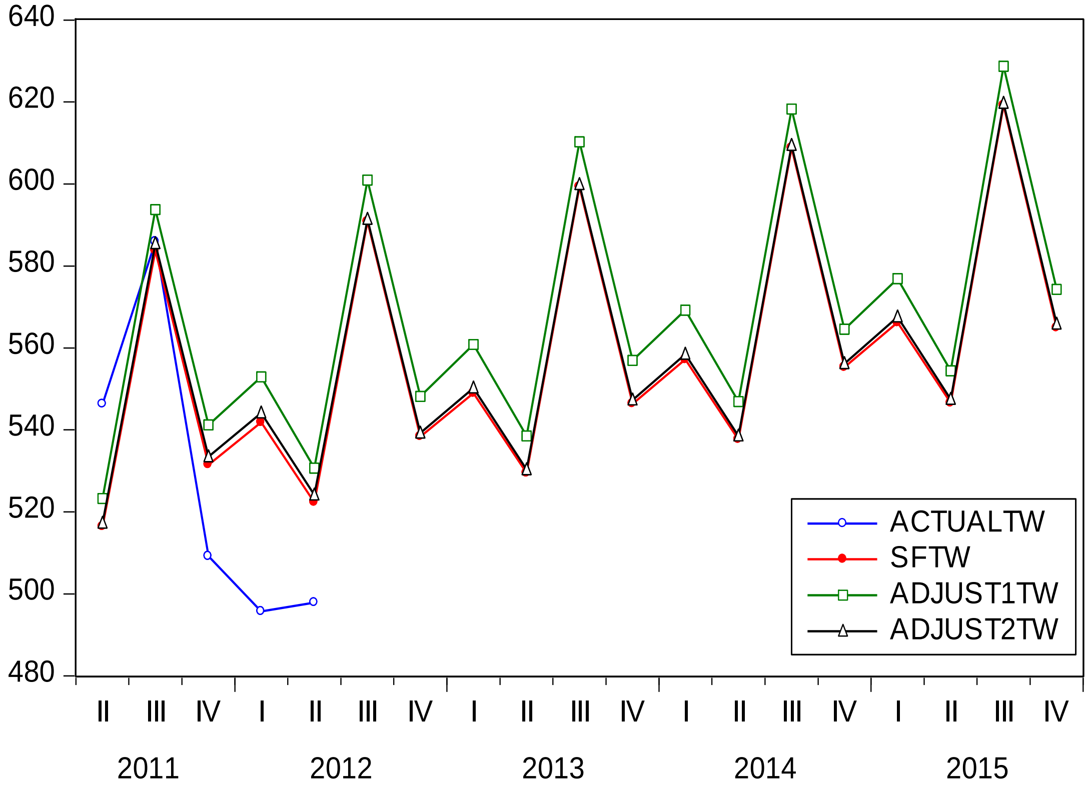
(a) Australia



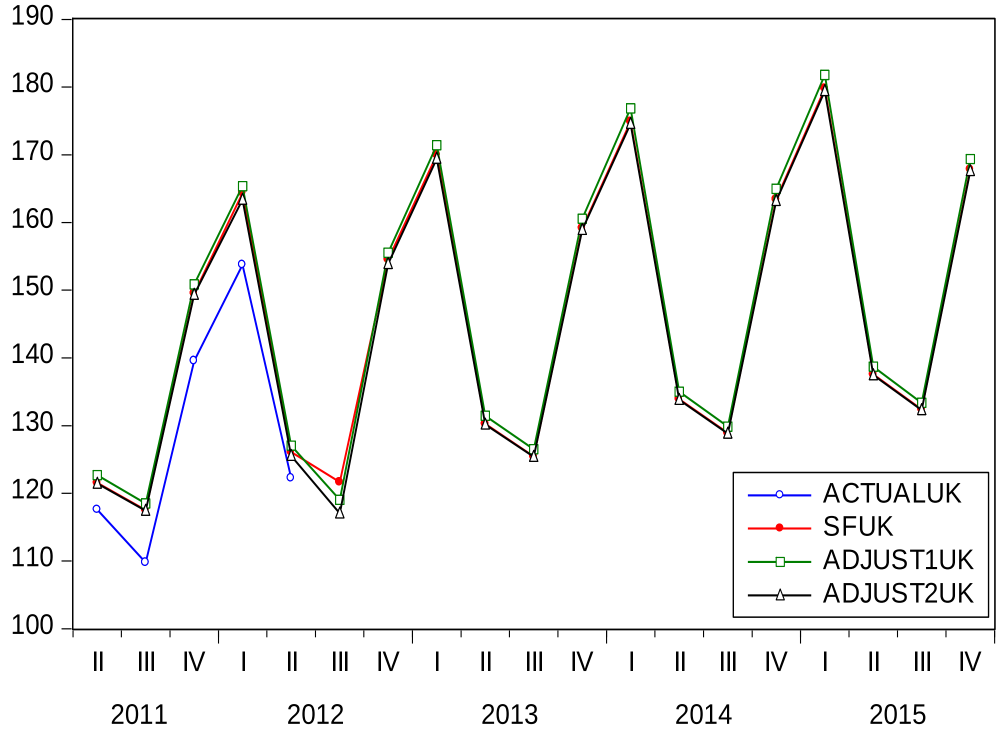
(b) China



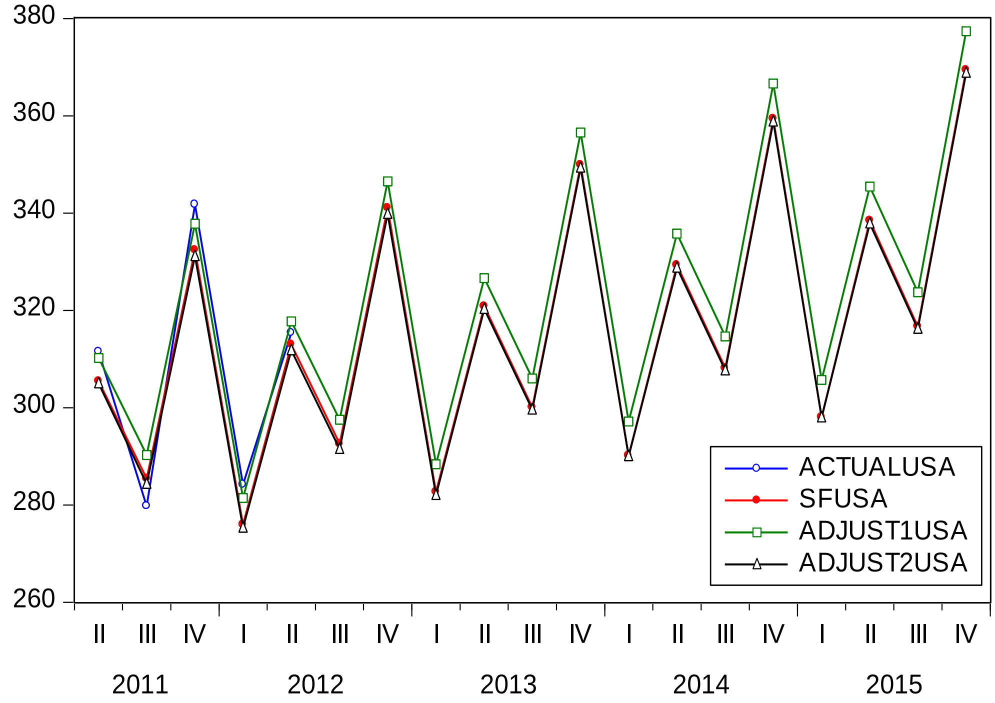
(c) Japan



(d) Taiwan



(e) UK



(f) USA

Figure 5.15 Comparison of actual arrivals and forecasts by market 2011Q2−2015Q4

Note: Actual: actual arrivals, SF: statistical forecasts, Adjust1: group forecasts in Round 1, Adjust2: group forecasts in Round 2.

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 5.19 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. The intercepts for three models (Australia, Taiwan, and the UK) were significantly less than zero, indicating that the forecasts for these markets were overestimated. The intercept for the China model was significantly greater than zero, suggesting that the forecasts for the Mainland market were underestimated. The above two findings are consistent with the visual judgment from the line plots in Figure 5.15. The intercept for the Japan model was negative but not significantly different from zero. Although the coefficient test result showed that the forecasts for this market were unbiased, more actual data points should have been included to confirm such a finding as it was probably due to the mixed impacts of underforecasting and overforecasting for individual quarters.

Generally, the experts consistently overestimated visitor arrivals for all of the markets except for the Mainland. 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' (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 optimistic 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) to the extent that the scenarios people generate in the context of making forecasts provide a mental script for how to behave.

H2b: Judgmentally adjusted forecasts are inefficient.

The previous section discussed testing for bias; the hypothesis above was formulated to examine the efficiency of forecasts. The most immediate information that a forecaster can bring into the forecasting adjustment is the time series data, the latest forecasts available to him/her, and the most recent forecast errors (Fildes et al., 2009). This hypothesis tests forecast efficiency by considering both the impacts of latest forecasts and forecast errors.

As shown in Table 5.19, β_0 was significantly different from zero at the 5% level for two rounds, indicating that the group forecasts were, on average, inefficient:

forecast errors in the experts' adjustments were correlated in adjacent quarters. A further examination into the forecasts of the individual experts showed the same result, namely that the efficiency condition was strongly rejected at the 5% significance level for experts' adjusted forecasts for the two rounds. One interpretation compatible with this finding is that the panellists systematically overestimated the permanence of past forecast errors when forecasting future visitor arrivals. In other words, the experts did not properly incorporate information about the time-series properties of visitor arrivals into their forecasts.

Furthermore, three Wald tests (as discussed in Section 3.7.3 of Chapter 3) were carried out to test the null hypothesis that group forecasts are weakly efficient. The last column of Table 5.20 presents the Wald test results. The efficiency condition was very strongly rejected at the 1% significance level for three sets of forecasts, namely the statistical forecasts and the group forecasts in the two Delphi rounds. In other words, the joint test of the assumptions that α_1 was significantly different from zero and β_1 was significantly different from unity was rejected, indicating that the three sets of forecasts were not conditionally efficient.

Table 5.19 also shows that the group forecasts for all six markets in the initial round were found to be inefficient (with β_0 significantly different from zero) but the forecasts for two markets (UK and USA) were efficient. Several studies have reached similar conclusions; for example, Lim and O'Connor (1995, 1996) found that the adjustment of initial judgmental forecasts based on a forecast received from another source (e.g. statistical forecasting method) tends to be inefficient. There are a number of reasons why forecasts based on managerial judgment (where the forecasts could be improved by modifying them to take into account information available to the forecaster at the time) are likely to be inefficient. Human cognitive limitations mean that people will struggle to optimally incorporate the effects of information from multiple sources into their forecasts (Fildes, 1991). These limitations may restrict human information processing capacity: people can only deal with information from one or two sources. Furthermore, when predicting these impacts, forecasters may overrely on the recall of single analogies from the past, and they may anchor too closely to these recalled effects (Lee, Goodwin, Fildes, Nikolopoulos, & Lawrence, 2007). In addition, the "escalation of commitment" literature demonstrates a strong reluctance in human judgment to modify a view already held about the future (Staw, 1976, cited in Fildes et al., 2009).

In addition, one feature of adjustment that is immediately apparent from Table 5.20 is that revision activity by the panellists led to an overall increase in the forecasts. This is not to say that all forecasts were revised upward, but it may indicate some degree of optimistic bias by the panellists concerned since long-term market/economic trends would have been allowed for by the trend element in the model.

To sum up the findings, the evidence presented in this section suggests that judgmentally revised forecasts were, on average, unbiased but were consistently inefficient in that they failed to incorporate all of the information from their own past forecasts and forecast errors. Thus, the findings support *H2a* and *H2b*. This is consistent with the findings of Musso and Phillips (2002), who found that forecasts that are unbiased always suffer from inefficiency or could be of poor accuracy as they convey little information to predict the future. Given that experts' predictions are biased and inefficient, 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.

Table 5.19 Regression coefficients for bias and inefficiency (Dependent variable: PE_t)

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

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

Table 5.20 Regression results (Dependent variable: Actual arrivals)

Model	Independent variable	Coefficient	t	F	R^2	Adjusted R^2	Wald test
1	Constant	-112.195	-1.906	2780**	0.990	0.9897	272.211**
	SF	1.399	52.727**				
2	Constant	-72.277	-1.377	3454**	0.992	0.9917	153.248**
	GF1	1.239	58.777**				
3	Constant	-53.247	-1.063	3774**	0.993	0.9924	113.95**
	GF2	1.188	61.436**				

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

(3) Hypotheses on the Delphi process

H3a: Forecast accuracy improves via the Delphi approach: Final Delphi forecasts are more accurate than the average of the initial estimates of the group members (i.e. statistical group).

As already discussed in testing hypothesis *H1a*, forecast accuracy improved over rounds in MAPE and RMSPE. To provide more evidence in testing *H3a*, a series of statistical tests were carried out using three error measures (APE, MAPE, and RSMPE) to examine the group experts' forecasting performance and the individual experts' forecasting performance. Regression analysis was applied to gain additional insights into the relative performance of the forecasts in Rounds 1 and 2.

The one sample *t*-test results in Table 5.21 show that the experts had significantly lower MAPE (t (14) = -3.302, p = 0.0025 < 0.01) and lower RMSPE (t (14) = -3.379, p = 0.0028 < 0.01) in the second round. This was further confirmed by a Wilcoxon signed-rank test with the null that the median of gapmape/gaprmspe equals zero (see Table 5.21). Furthermore, accuracy also increased over rounds for MAPE, RMSPE, and the U statistic irrespective of which consensus measure was used (see Table 5.11). The results from the above two tests showed that the performance of the group panellists significantly improved as a result of using the Delphi technique.

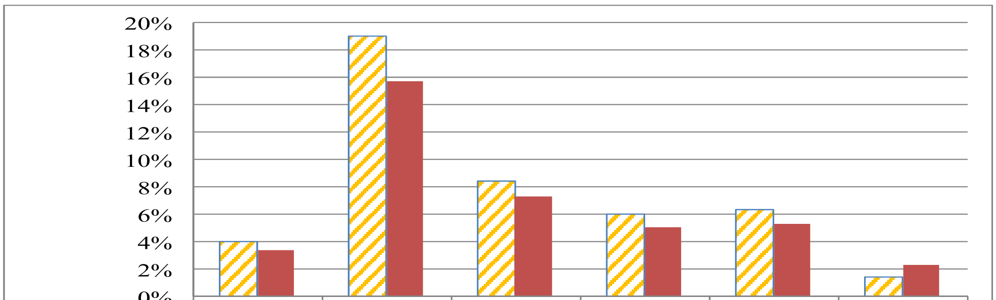
Table 5.21 Results for one-sample t test and Wilcoxon signed rank test

Test	One sample *t* test		Wilcoxon signed rank test	
	t	*p*. (1-tailed)	*z*	*p*. (1-tailed)
gapmape	-3.302	.003***	-1.500	0.071*
gaprmspe	-3.379	.003***	-2.637	0.003***

Note: ***, ** and * indicate significance at the 1%, 5% and 10% level, respectively.

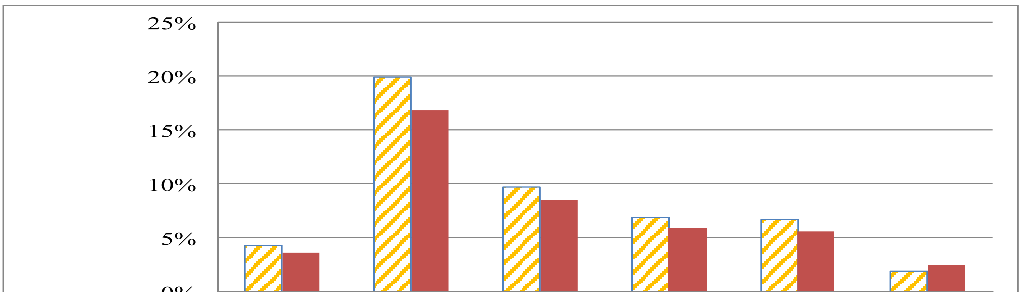
Even though forecast adjustments improved forecast accuracy on average, the level of improvement varied across source markets for the evaluation period 2011Q2−2012Q2. As shown in Table 5.12, the greatest mean improvement in accuracy over the two rounds was achieved in the UK market (with a 17.67% reduction in MAPE), followed by the Mainland and Taiwan. During the evaluation period 2011Q2−2012Q2, the experts' adjustments from Round 1 to Round 2 in five out of the six markets achieved an improvement in MAPE and RMSPE; however, for the USA, accuracy deteriorated after making adjustments for both rounds.

The bar graphs in Figures 5.16 to 5.18 compare the values of MAPE, RMSPE, and the U statistic among the six source markets from Round 1 to Round 2. According to these three bar graphs, forecast accuracy was, on average, improved over rounds for five out of the six source markets (Australia, China, Japan, Taiwan, and the UK) but decreased for the USA market.



	Australia	China	Japan	Taiwan	UK	USA
MAPE(R1)	4.03%	19.04%	8.39%	6.00%	6.37%	1.41%
MAPE(R2)	3.38%	15.69%	7.26%	5.05%	5.24%	2.22%

Figure 5.16 MAPE by market 2011Q2−2012Q2



	Australia	China	Japan	Taiwan	UK	USA
RMSPE (R1)	4.32%	19.98%	9.74%	6.87%	6.63%	1.85%
RMSPE (R2)	3.57%	16.83%	8.47%	5.91%	5.57%	2.36%

Figure 5.17 RMSPE by market 2011Q2−2012Q2

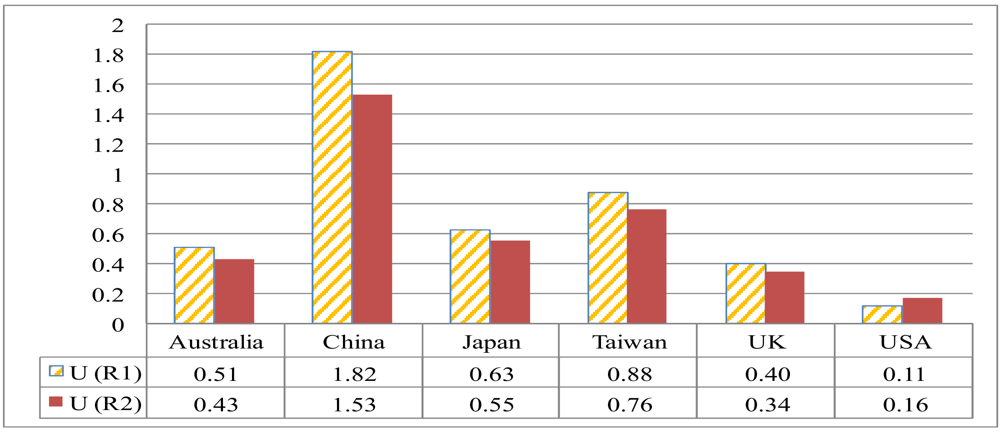


Figure 5.18 Theil's U statistic by market 2011Q2−2012Q2

More detailed analyses of the performance statistics based on APE are provided in Table 5.17. The results in this table again confirm that the group forecasts from the second round were more accurate than those from the first round. Improvements in forecast accuracy over rounds were observed for four of the five quarters in the cases of Australia, Japan, and Taiwan, for all quarters in the case of China, and for one quarter in the case of the UK.

As shown in Section 3.7.2 of Chapter 3, the forecasting performance of two rounds was also evaluated by conducting regression analysis. Table 5.20 shows that a slightly higher R^2 was obtained with the adjusted forecasts in the second round as the independent variable, but the difference was relatively small. Table 5.20 also illustrates that when the second round's group forecast was used as a predictor, the intercept of the regression line was closer to zero and the slope was closer to unity. Taken collectively the above findings suggest that the second round group forecasts would probably serve as a better predictor of actual arrivals.

The analysis presented earlier in this section mainly concerned the group forecasts (or using the mean forecasts as the consensus forecasts). To provide a holistic view of the evaluation results, the forecasting performance of individual experts should also be evaluated. The results in Table 5.22 show that over half

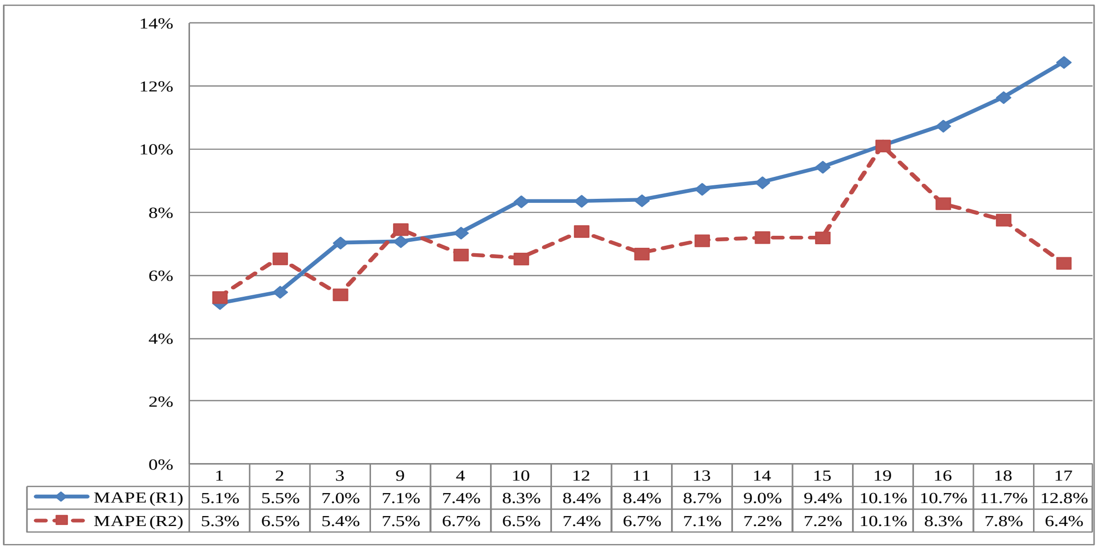
(72.2%) of the experts' forecasts obtained a MAPE lower than 10 per cent in the first round and an overwhelming 94.1 per cent of experts' forecasts had an MAPE lower than 10 per cent in the second round. Among the experts who ranked in the top five, four worked in the tourism industry.

Table 5.22 Forecast accuracy summary by individual expert

Group	ID (Ranking)	MAPE (R1)	MAPE (R2)	RMSPE (R1)	RMPSE (R2)	Mean MAPE	Mean RMSPE
Industry	1	5.12	5.31	6.64	6.71	5.21	6.68
Academic	2	5.48	6.54	7.04	9.18	6.01	8.11
Industry	3	7.04	5.39	10.82	7.49	6.22	9.16
Industry	4	7.35	6.66	10.84	9.32	7.01	10.08
Industry	5	7.03	–	8.71	–	7.03	8.71
Academic	6	–	7.21	–	9.87	7.21	9.87
Industry	7	–	7.21	–	9.87	7.21	9.87
Industry	8	7.22	–	10.14	–	7.22	10.14
Academic	9	7.08	7.47	9.47	10.07	7.27	9.77
Academic	10	8.35	6.53	10.83	8.69	7.44	9.76
Academic	11	8.38	6.69	12.86	8.81	7.54	10.83
Academic	12	8.36	7.40	12.04	9.98	7.88	11.01
Academic	13	8.74	7.11	12.70	9.64	7.93	11.17
Industry	14	8.95	7.21	10.80	9.87	8.08	10.34
Industry	15	9.45	7.20	13.41	9.49	8.32	11.45
Academic	16	10.75	8.29	12.79	11.18	9.52	11.99
Academic	17	12.77	6.39	16.00	8.59	9.58	12.30
Academic	18	11.65	7.76	16.51	10.06	9.71	13.29
Academic	19	10.12	10.12	19.06	19.06	10.12	19.06
Academic	20	13.04	–	14.48	–	13.04	14.48
Industry		**7.45**	**6.50**	**10.20**	**8.79**	**7.04**	**9.55**
Academic		**9.52**	**7.41**	**13.07**	**10.47**	**8.60**	**11.80**
Average		**8.72**	**7.09**	**11.95**	**9.88**	**7.98**	**10.90**

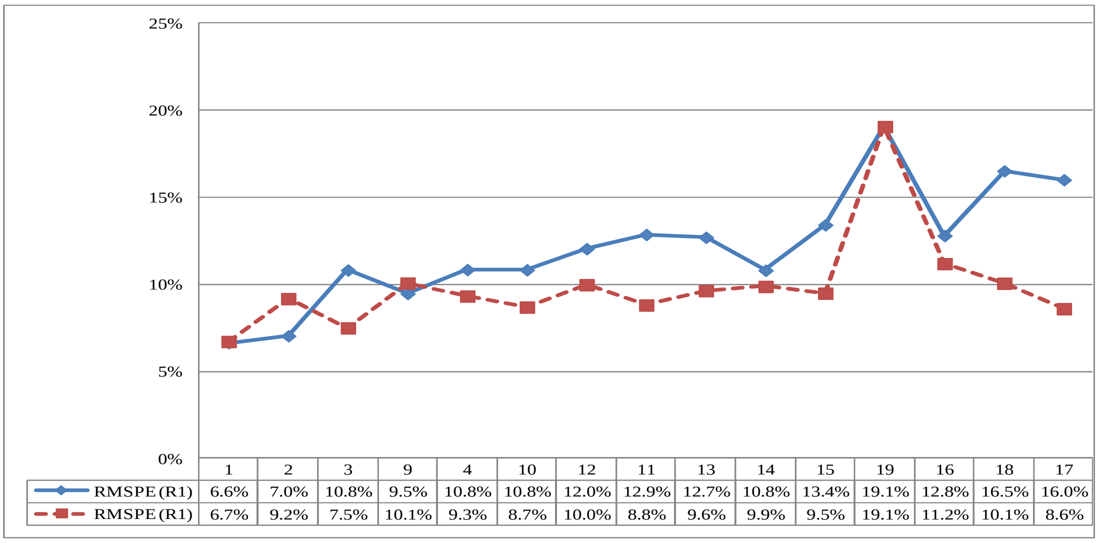
To examine the forecasting performance of the 15 individual experts who took part in both rounds, MAPE and RMSPE were calculated, and they are plotted in Figures 5.19 and 5.20 respectively. It is clearly shown that individual experts improved their performance over rounds. To examine whether there was statistical significance in such performance difference, the Wilcoxon signed-rank test was employed. It was found that the 15 experts had significantly lower MAPE (z = -2.783, p = 0.002 < 0.01) and lower RMSPE (z = -2.668, p = 0.003 < 0.01) in the second round. This finding confirmed that the performance of individual experts was significantly improved by utilizing the Delphi approach. These results indicate that either the experts may have found Delphi a useful tool for pressuring them to provide better forecasts or the presence of the new information available in the second round enabled them to make their forecasts more accurate.

Overall, the findings discussed earlier in this section support the hypothesis (*H3a*) that the use of the Delphi technique could provide significantly better forecasts than the average of group members' initial judgments. These results are consistent with the rationale underlying the Delphi technique.



	1	2	3	9	4	10	12	11	13	14	15	19	16	18	17
MAPE (R1)	5.1%	5.5%	7.0%	7.1%	7.4%	8.3%	8.4%	8.4%	8.7%	9.0%	9.4%	10.1%	10.7%	11.7%	12.8%
MAPE (R2)	5.3%	6.5%	5.4%	7.5%	6.7%	6.5%	7.4%	6.7%	7.1%	7.2%	7.2%	10.1%	8.3%	7.8%	6.4%

Figure 5.19 MAPE by 15 individual experts 2011Q2-2012Q2



	1	2	3	9	4	10	12	11	13	14	15	19	16	18	17
RMSPE (R1)	6.6%	7.0%	10.8%	9.5%	10.8%	10.8%	12.0%	12.9%	12.7%	10.8%	13.4%	19.1%	12.8%	16.5%	16.0%
RMSPE (R1)	6.7%	9.2%	7.5%	10.1%	9.3%	8.7%	10.0%	8.8%	9.6%	9.9%	9.5%	19.1%	11.2%	10.1%	8.6%

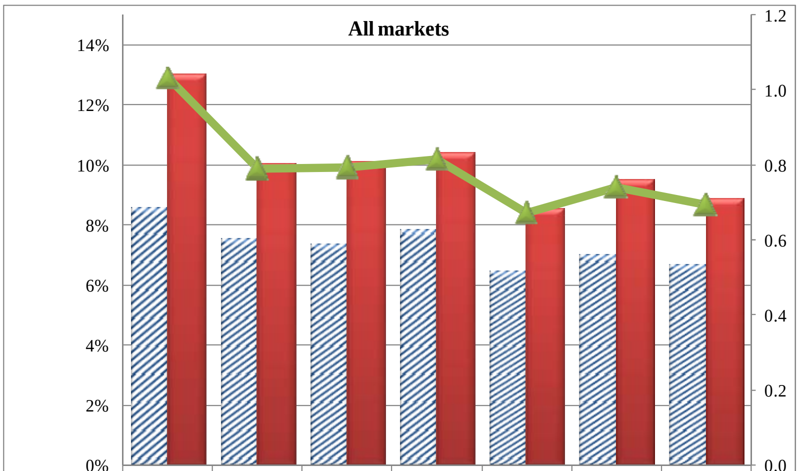
Figure 5.20 RMSPE by 15 individual experts 2011Q2-2012Q2

H3b: Combining experts' forecasts with the mean is more accurate than other consensus measures.

In this study, consensus forecasts were made by averaging individual experts' forecasts in each round. In the existing Delphi literature, as discussed in Section 2.4.2, median is another commonly used measure to reach consensus forecasts. Thus, it is interesting to compare the forecasting performance of mean and median group forecasts as this will probably provide suggestions for future Delphi studies in forecasting tourism demand.

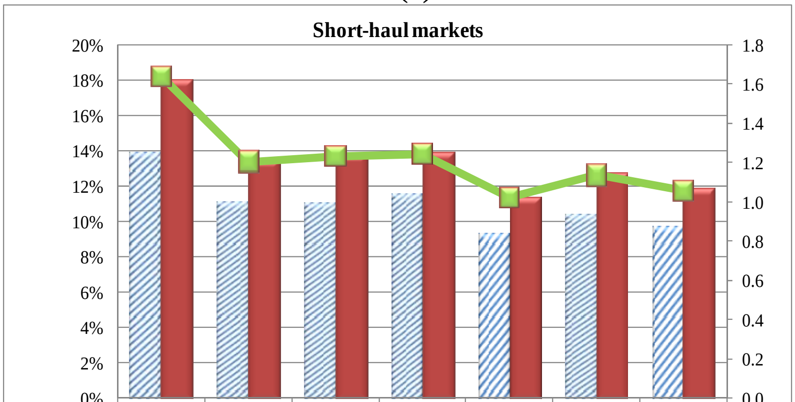
At the end of the first iteration, the average of the MAPE for all adjusted forecasts combined was 7.36 per cent (median) compared to a mean of 7.54 per cent. The results obtained by RMSPE as a measure of accuracy were different from the ones obtained from MAPE: 11.51% (median) and 10.12% (mean). For the individual quarters, the median forecasts were more accurate than the mean forecasts according to the MAPE across all forecasting horizons and the RMSPE from the 3- to 5-quarter-ahead forecasts (see Table 5.13).

In the second round, the performance of the mean as a consensus estimate was much better compared to the median for MAPE and RMPSE. In terms of the performance for single forecasting horizons, the same results were observed. As shown in Figure 5.21, the accuracy of the mean response of the final group forecast was much higher than either the median or the self-rating weighted mean.



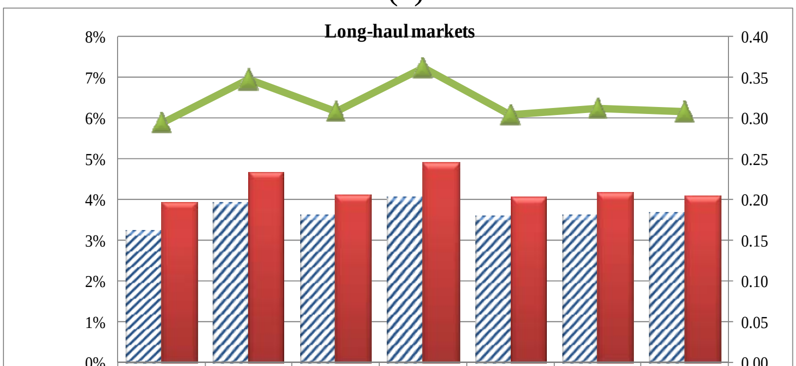
	SF	Mean(R1)	Median(R1)	Smean(R1)	Mean(R2)	Median(R2)	Smean(R2)
MAPE	8.59%	7.54%	7.36%	7.81%	6.47%	7.03%	6.71%
RMSPE	13.04%	10.06%	10.14%	10.42%	8.56%	9.53%	8.89%
Theil's U	1.03	0.79	0.79	0.81	0.67	0.74	0.69

(a)



	SF	Mean(R1)	Median(R1)	Smean(R1)	Mean(R2)	Median(R2)	Smean(R2)
MAPE	13.92%	11.14%	11.07%	11.56%	9.33%	10.42%	9.73%
RMSPE	18.02%	13.43%	13.74%	13.90%	11.40%	12.81%	11.88%
Theil's U	1.64	1.20	1.23	1.24	1.02	1.14	1.06

(b)



	SF	Mean(R1)	Median(R1)	Smean(R1)	Mean(R2)	Median(R2)	Smean(R2)
MAPE	3.25%	3.94%	3.64%	4.07%	3.61%	3.63%	3.68%
RMSPE	3.93%	4.69%	4.12%	4.91%	4.05%	4.17%	4.10%
Theil's U	0.29	0.35	0.31	0.36	0.30	0.31	0.31

(c)

Figure 5.21 Comparisons by consensus measure

Note: SF: statistical forecasts, Mean: mean forecasts of experts' adjustments, Median: median forecasts of experts' adjustments, Smean: mean forecasts weighted by self-rated expertise.

To sum up, a reasonable consensus estimate of the group judgments would be the simple average or the mean of individual experts' forecasts. The median is preferable to the mean as a measure of central tendency in the initial round of Delphi. However, to obtain a more accurate final group forecast, the mean with equal weight for all individual experts' estimates is recommended. The above findings suggest that there is no conclusive evidence to support hypothesis *H3b*.

(4) Hypotheses on self-rated expertise

H4a：Higher self-rated expertise is related to more accurate Round 1 forecasts.

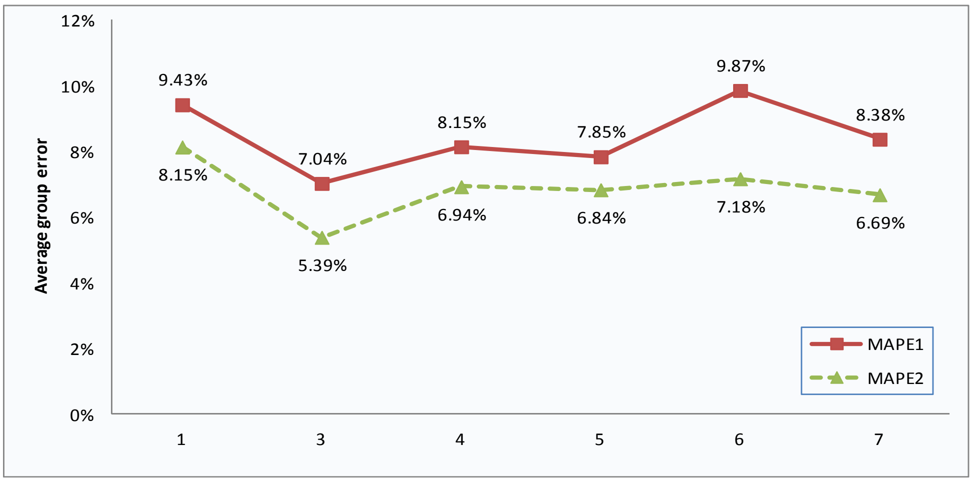
At the beginning of the main Delphi survey, the panellists were required to rank their own expertise in tourism demand forecasting on a 7-point scale. To test *H4a* and *H4b*, analyses contrasting higher self-rated expertise scores with those on the less-expert side were conducted to examine whether there were statistically significant differences in accuracy between those subgroups.

Experts with different self-rated expertise were categorized into two groups: low (scores ranging from 1–4) and high (scores ranging from 5−7). In Round 1, the low-expertise group consisted of five experts, while the high-expertise group had 13 experts. On average, self-rated expertise appeared to discriminate in terms of the accuracy of the individual experts. However, the experts with higher self-rated expertise did not provide significantly more accurate forecasts than those with lower expertise scores. This was confirmed by the Mann-Whitney tests, which showed that these differences did not reach statistical significance (see Table 5.23).

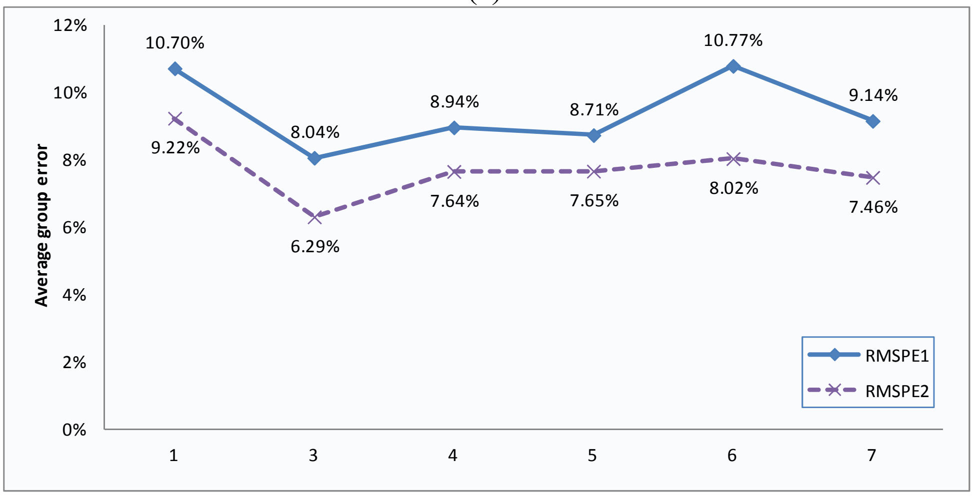
Table 5.23 Results of Mann-Whitney tests for comparisons between group expertise

Round		Self-rated expertise	N	Mean Rank	Z	p (1-tailed)
R1	MAPE	Low	5	9.60	-0.049	0.500
		High	13	9.46		
R2	MAPE	Low	5	8.20	-0.122	0.477
		High	10	7.90		
R1	RMSPE	Low	5	10.20	-0.345	0.387
		High	13	9.23		
R2	RMSPE	Low	5	8.60	-0.367	0.384
		High	10	7.70		

Figure 5.22 plots the relationship between group self-rating scores and the mean group errors over rounds. In contrast to Dalkey's (1969) findings, the current study did not show a clear inverse relationship between group self-rating scores and group errors. However, Figure 5.22 provides visual evidence that it is likely that the higher the average confidence rating of group experts, the smaller the group errors. To provide a better understanding of the different findings, it is also worth pointing out the difference between Dalkey's study and the current study in terms of the self-rated expertise data: Dalkey's (1969) study asked respondents to give a confidence rating for each question, whereas this study only asked the panellists to give an overall rating of their expertise in tourism demand forecasting.



(a)



(b)

Figure 5.22 MAPE and RMSPE of average group self-rating scores

Question: Self rating of expertise in tourism demand forecasting
Very little = 1 ———— 7 = Excellent
Your option: ☐ (Please input a number between 1 and 7 here!)

Additional insights into the accuracy among group self-ratings were obtained by conducting two correlation analyses for the academic and industry expert groups. The results in Table 5.24 show that in Round 1, the average group MAPE (or RMPSE) and the self-rated expertise were negatively correlated for both two groups. However, such an inverse relationship was insignificant at any significance level, suggesting that self-rated expertise may not be efficient as a predictor of initial accuracy.

In Round 2, the correlation results for the two expert groups were different. The correlation between self-rated expertise and group errors reached high significance in MAPE (r = -0.99, p = 0.01) and RMSPE (r = -0.98, p = 0.01) for the academic experts. In contrast, this relationship became positive for the industry group experts irrespective of the error measures used.

Table 5.24 Relationship of average group self-rated expertise and accuracy

Group	Self-rating	MAPE (%)		RMSPE (%)	
		R1	R2	R1	R2
Industry	3	7.04	5.39	8.44	6.46
	4	8.15	6.94	9.21	7.92
	5	7.20	6.25	8.24	7.31
	6	7.22	7.21	8.25	8.22
Pearson Correlation		**-0.10**	0.76	**-0.43**	0.78
p. (1-tailed)		0.45	0.12	0.29	0.11
Academic	1	9.43	8.15	11.44	9.74
	5	8.50	7.23	9.76	8.21
	6	10.40	7.17	11.53	8.26
	7	8.38	6.69	9.67	7.65
Pearson Correlation		**-0.19**	**-0.99**	**-0.53**	**-0.98**
p. (1-tailed)		0.41	0.01	0.23	0.01

The above analyses suggest that the experts with higher self-rated expertise were likely to produce more accurate forecasts in Round 1 when the experts were split into industry and academic groups, which supports hypothesis *H4a*. However, after iteration, this inverse relationship became unclear. These findings are partly different from those of Rowe and Wright (1996), who concluded that the relationship between self-rated expertise and accuracy was less clear after the implementation of a structured group procedure, but self-rated expertise provided a predictor of initial accuracy.

Some possible explanations for the insignificant correlation between self-rated expertise and group errors (for MAPE and RMSPE) in the first round are provided. First, the overall self-rated expertise in tourism demand forecasting reported by some experts may not have truly reflected these experts' objective forecasting expertise.

Second, the expertise scores were not equally distributed from one to seven, and this might have caused the biased results of the correlation analysis. Table 5.24 shows that the industry experts claimed that their self-rated expertise ranged between 3 and 6, while most of the academic experts rated their forecasting expertise between 5 and 6. Dalkey, Brown, and Cochran (1969) concluded that group size and average group error are closely related − reducing the size of the group would result in a substantial reduction in average accuracy. They further illustrated that the size effect could then mask any improvement resulting from the self-rating effect. Therefore, to increase accuracy, the size of subgroups should be substantial for both the higher and lower self-rating subgroups.

Last but not least, the 7-point self-rating scale may not have adequately reflected the real expertise of the panellists. The reason for this is perhaps that the experiences of those experts were noncomparable or that no objective measurements of past performance exist. In addition, some of the experts selectively chose some but not all of the six markets to make their adjustments. Even though they may have obtained a higher forecast accuracy for the markets they selected, their forecasting performance might have deteriorated due to those markets they did not select, and vice versa.

H4b：High self-rated expertise is related to a low propensity to make judgment changes over rounds.

The correlation analysis reported in Table 5.25 shows that, on average, there was a significant relationship between self-rated expertise and the percentage changes from the first round to the second round, $r = -0.16$, $p < 0.01$. It was found that the higher the self-rated expertise, the less changes the panellist made over rounds, which supports hypothesis *H4b*. It was the same for all of the individual markets except for the Mainland. This finding indicates that experts with higher confidence may have been less willing to revise their initial judgments and consequently drew other members of the group towards their own points of view even in the absence of interpersonal interactions, which is consistent with the findings of Larréché and Moinpour (1983).

Table 5.25 Relations of percentage changes over rounds and self-rated expertise

	Australia	China	Japan	Taiwan	UK	USA	Overall
Pearson Correlation	-0.091*	-0.006	-0.254***	-0.087*	-0.267***	-0.415***	-0.165***
p (1-tailed)	0.063	0.458	0.000	0.071	0.000	0.000	0.003
N	285	285	285	285	285	285	285

(5) Hypotheses on the characteristics of judgmental forecasting tasks

H5a: Forecast accuracy decreases with data variability.

The main purpose of the hypothesis test described in this section was to examine the relationship between the variability of a time series (as measured by the CV of the raw data) and the relative accuracy of statistical and judgmentally adjusted forecasts. Based upon the CV values, the six series were segmented into three categories: low variability (below 0.3), medium variability (between 0.3 and 0.5), and high variability (above 0.5). Table 5.17 shows the CV values for each time series (i.e. arrivals for individual markets). To illustrate the initial judgment on the impact of data variability on forecast accuracy, Figure 5.23 shows the error bars which represent the APE of three sets of forecasts averaged by data category.

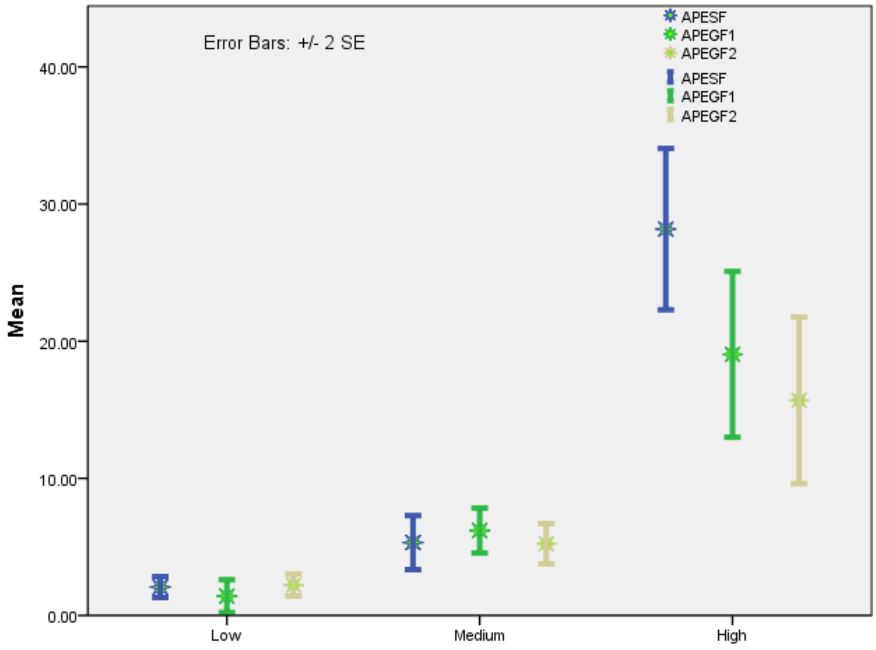


Figure 5.23 Error bar charts in APE

As the analysis on visually detecting the pattern of the error bars earlier only provided an initial judgment on the relationship between accuracy and data variability, the Kruskal-Wallis test was applied to check its statistical significance. Table 5.26 shows that the accuracy evaluated by APE was significantly related to the data variability as the *p* values in the Kruskal-Wallis test were less than 0.05. Mann-Whitney tests were used to follow up this finding. A Bonferroni correction was applied and so all effects were reported at a 0.017 level of significance.

Significant accuracy differences were found between the low- and high-variability groups, and the medium- and high-variability groups in all of the comparisons shown in Table 5.26. Only one insignificant result was found when evaluating the accuracy of statistical forecasts between the low- and medium-variability groups: the *p* value of the Mann-Whitney U statistic was greater than the significance level of 0.017.

Table 5.26 Comparison of the forecast accuracy of different volatility data groups

Dependent Variable	**Kruskal-Wallis test**		**Mann-Whitney test**					
	H statistic	*p*	Multiple Comparisons		*U*	*z*	*p*	Effect size (*r*)
APE$_{SF}$	14.62	0.001	Low (7.40)	Medium (14.40)	22.00	-1.90	0.030 (>0.017)	-0.38 (M)
			Low (7.40)	High (28.00)		-2.61	**0.008** (<0.017)	-0.83 (L)
			Medium (14.40)	High (28.00)		-3.40	**0.000** (<0.017)	-0.68 (L)
J-T test			*J* statistic		203	*z*	3.90	0.71 (L)
APE$_{GF1}$	17.97	0.000	Low (3.80)	Medium (15.45)	4.00	-3.13	**0.000** (<0.017)	-0.63 (L)
			Low (3.80)	High (27.40)		-2.61	**0.008**(<0.017)	-0.83 (L)
			Medium (15.45)	High (27.40)		-3.19	**0.000** (<0.017)	-0.64 (L)
J-T test			*J* statistic		218	*z*	4.54	0.83 (L)
APE$_{GF2}$	14.61	0.000	Low (6.40)	Medium (14.80)	17.00	-2.24	**0.012** (<0.017)	-0.45 (M)
			Low (6.40)	High (27.40)		-2.61	**0.009** (<0.017)	-0.83 (L)
			Medium (14.80)	High (27.40)		-3.19	**0.000** (<0.017)	-0.64 (L)
J-T test			*J* statistic		205	*z*	3.98	0.73 (L)

Note: Figures in parentheses are mean ranks.

Furthermore, Jonckheere's tests revealed a significant trend in the data: the sign of *z*-value was positive for three sets of comparisons (see Table 5.26), which means that as the level of data variability increased, the median APE increased, or the forecast accuracy decreased. This finding supports *H5a*, which stated that a higher level of data variability could be associated with lower forecast accuracy.

The last column of Table 5.26 reports the effect size of each test; for example, a value of -0.38 represents a medium effect for the Mann-Whitney test, suggesting that the effect accounted for 38 per cent of the total variance.

The nonparametric tests above were used to calculate and compare the difference in the median values for different groups. The Hochberg's GT2 post-hoc test (used in case of varying sample sizes) was also employed to check the mean difference in accuracy for the three levels of data categories. It was found that there was a significant effect of data variability on the accuracy of all three sets of forecasts as the *F* statistics were all less than 0.05 (see Table 5.27). The results from multiple comparisons showed that forecast accuracy measured by APE reduced with the increasing level of data variability. Specifically, data of lower variability were related to higher forecast accuracy or lower APE value. This again confirmed the findings from the Jonckheere's test results.

Table 5.27 Comparison of forecast accuracy based upon level of data variability by one-way ANOVA

Dependent Variable	Levene Statistic	F	Multiple Comparisons		Mean Difference	*p.* (1-tailed)
APE_{SF}	1.87	58.11**	Low	Medium	-3.25	0.20
			Low	High	-26.10	0.00
			Medium	High	-22.86	0.00
APE_{GF1}	2.02	26.83**	Low	Medium	-4.79	0.04
			Low	High	-17.63	0.00
			Medium	High	-12.84	0.00
APE_{GF2}	2.42	18.82**	Low	Medium	-3.01	0.16
			Low	High	-13.48	0.00
			Medium	High	-10.46	0.00

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

H5b: Experts' intervention is more valuable for the more variable time series.

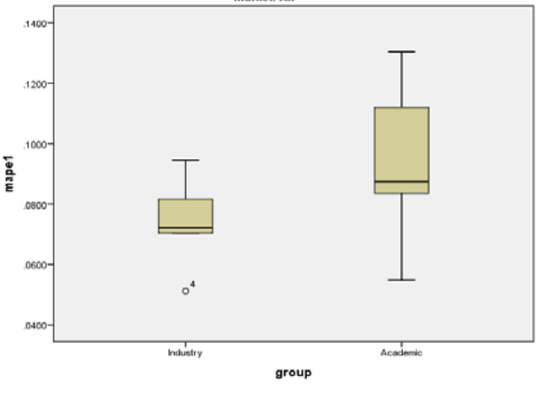
It can be seen from Figure 5.23 that as the arrivals series became more variable, the advantage of expert intervention became more evident; on average, forecast accuracy increased with the level of data variability. Figure 5.23 also shows that according to APE, experts' adjustment did not always improve forecast accuracy for all categories of arrivals series. For example, the experts' initial adjusted forecasts were slightly better than the original statistical forecasts for the low-variability series; however, the experts' final adjusted forecasts were not significantly more accurate than either the statistical forecasts or the Round 1 forecasts. For the high-variability series, it was evident that the forecasts after adjustment were much more accurate than the unadjusted forecasts (i.e. statistical forecasts). This implies that it could be difficult for statistical methods to achieve reasonable forecasts when the data are more variable, whereas adjusted forecasts reinforced by experts' domain knowledge could yield relatively better forecasts. In other words, the statistical forecasting model is more consistent and is a better performer with stable data; on the other hand, experts' domain knowledge becomes very valuable for accuracy improvements in variable series where there is more apparent randomness that cannot be captured fully by statistical models (Sanders & Ritzman, 1992). In short, the above analysis supports hypothesis *H5b* that experts' judgments are more valuable for the more variable series in terms of forecasting arrivals.

Sanders and Ritzman (1992) drew a similar conclusion, namely that when a series has low coefficients of variation, statistical time-series methods outperform judgmental forecasters who have expertise relating to the variables to be forecast. They also found that as the volatility of a series increases, experts increasingly outperform time-series forecasting methods. One plausible explanation of why experts performed worse in stable series is that they tend to overreact to noise (or fluctuations) in the time series, particularly in conditions of high noise (Goodwin & Fildes, 1999; Sanders & Ritzman, 1992). As suggested by Willemain (1989), the most significant background variable in forecasting a series is its inherent difficulty. If a series is easy to forecast, then the initial automatic statistical forecasts are likely to be quite accurate, leaving little room for improvement.

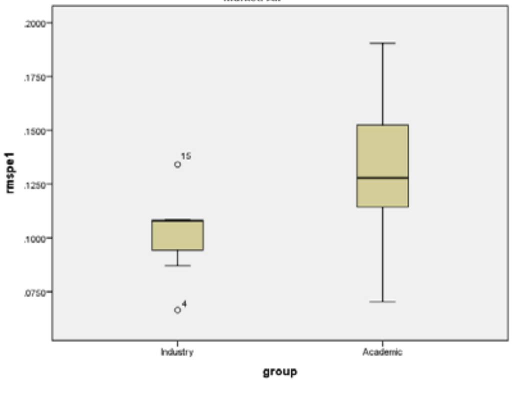
In forecasting arrivals in this study, different levels of data variability were found among the different markets. Specifically, lower data variability was related to relatively more mature markets, while higher data variability was associated with more volatile and vibrant markets. The findings discussed earlier in this section suggest that when the market is unstable, there may be greater potential for effective human intervention since the forecaster usually has a great deal more information available to him/her than is utilized in the forecasting model. This is consistent with the findings of Diamantopoulos and Mathews (1989). Managerial information is typically qualitative and incomplete. It comes from many sources and is difficult to incorporate into a quantitative forecasting model, and therefore the effectiveness of judgmental revisions could vary between source markets facing different social, economic, and environmental conditions; so some markets will be easier to assess/predict judgmentally than others.

H5c: Judgmentally adjusted forecasts from experts with more contextual knowledge are more accurate than those from experts with less contextual knowledge.

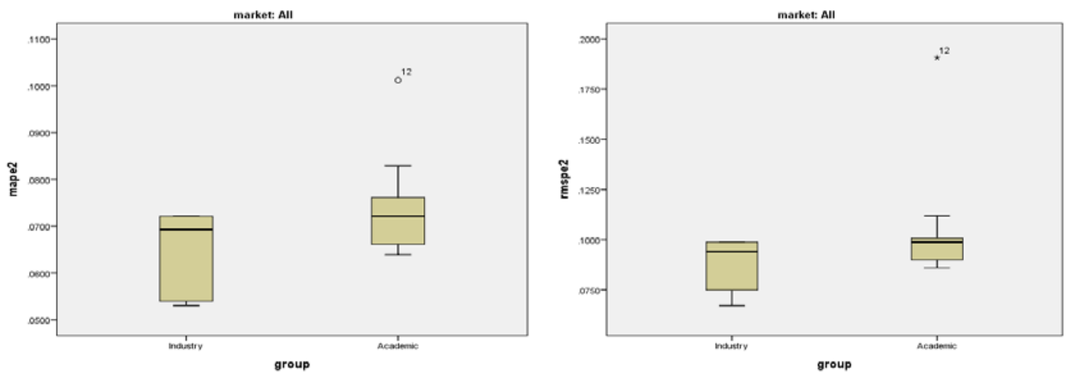
The results of the exploratory analysis are shown in Figures 5.24 and 5.25. The boxplots show the median values for MAPE and RMSPE between the two expert groups over rounds, while the error bars show the mean values along with the 95% confidence intervals. An important difference can be observed between the industry and academic experts: the median and mean values of MAPE and RMSPE for the industry experts were lower than those of the academic experts, providing visual evidence that the industry experts produced more accurate forecasts than the academic experts.



(a) R1: MAPE

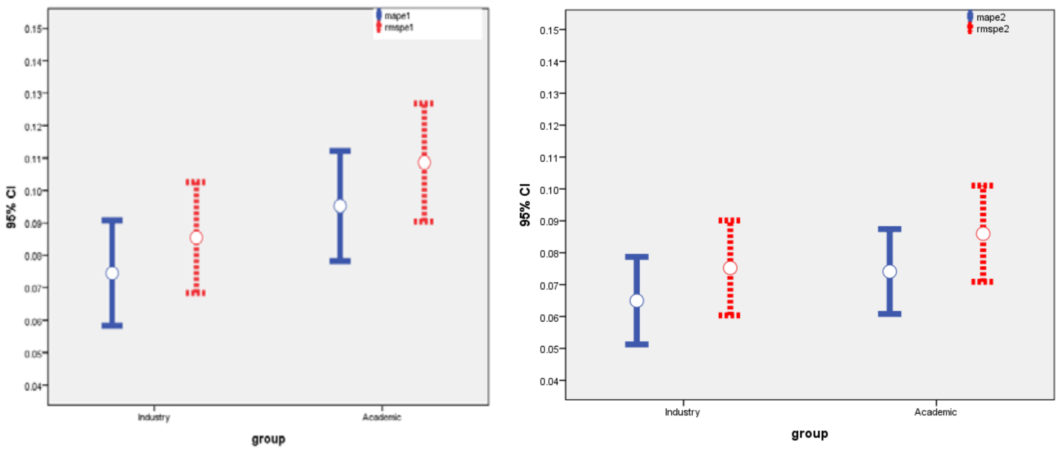


(b) R1: RMSPE



(c) R2: MAPE (d) R2: RMSPE

Figure 5.24 Boxplots of MAPE and RMSPE by rounds and expert group



(a) R1 (b) R2

Figure 5.25 Error bars of MAPE and RMSPE by rounds and expert group

In addition to the comparison of accuracy measures, statistical tests were conducted to examine whether there was any significant accuracy difference between the industry and academic experts. One finding that emerged from the Mann-Whitney tests was that the forecasts made by the industry practitioners were significantly better than those made by the academic researchers based on MAPE and RMSPE (see Table 5.28). Specifically, Table 5.28 shows that the industry experts' forecasts were significantly better than academic experts' forecasts at the 5% level in the initial round of the Delphi survey. In Round 2, the accuracy difference became less clear but was still significant at the 10% level. This result supports the statement in *H5c* that experts with more contextual knowledge make forecasts that are more accurate. This is consistent with past research findings that the incorporation of forecasters' contextual knowledge into statistical forecasts can be helpful in improving the forecast accuracy (e.g. Sanders & Ritzman, 1992).

The effect size was also calculated accordingly after each test. As shown in Table 5.28, the effect respectively explained 19%, 21%, 14%, and 14% of the total variance in the four tests, suggesting a medium to large effect. This indicates that the effect of the significant difference between the two expert groups was a fairly substantive finding.

Table 5.28 Accuracy difference in two expert groups: The Mann Whitney tests

Test	**R1**		**R2**	
	MAPE	RMSPE	MAPE	RMSPE
Mann-Whitney U	18.00	17.00	13.00	13.00
Wilcoxon W	46.00	45.00	28.00	28.00
Z	-1.86	-1.95	-1.47	-1.47
p (1-tailed)	0.03	0.03	0.08	0.08
Effect size	-0.44 (M)	-0.46 (M)	-0.38 (M)	-0.38 (M)
% of variance	19%	21%	14%	14%

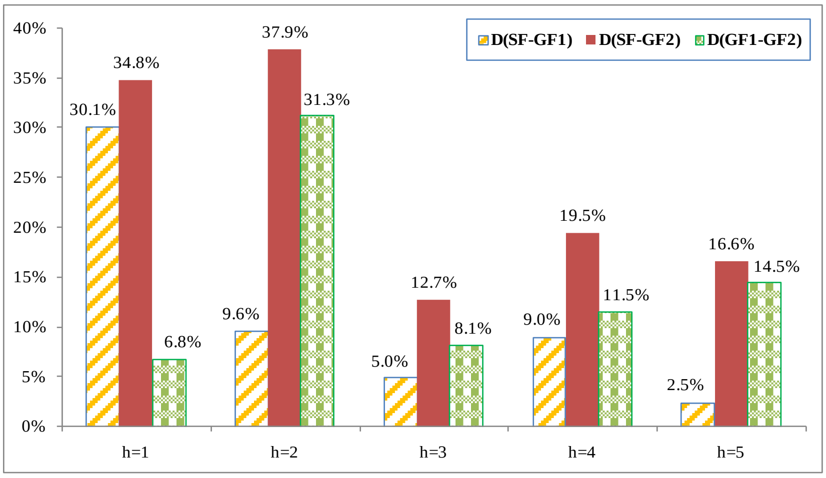
H5d: Improvement in accuracy from judgmental adjustments decreases over time.

As discussed in Section 3.6.5 of Chapter 3, most studies have agreed that the longer the time horizon, the less accurate the forecasts. The following analyses evaluated the accuracy of arrivals forecasts made in 2011Q2 or five quarters in advance. The forecasts covered a span of five-quarter forecasts from 2011Q2 to 2012Q2. Note that a current quarter (h = 1) forecast is actually a one-step-ahead forecast as the current-quarter value is unknown when the forecast is made; a next-quarter forecast is a two-step-ahead forecast, and so on up to h = 5.

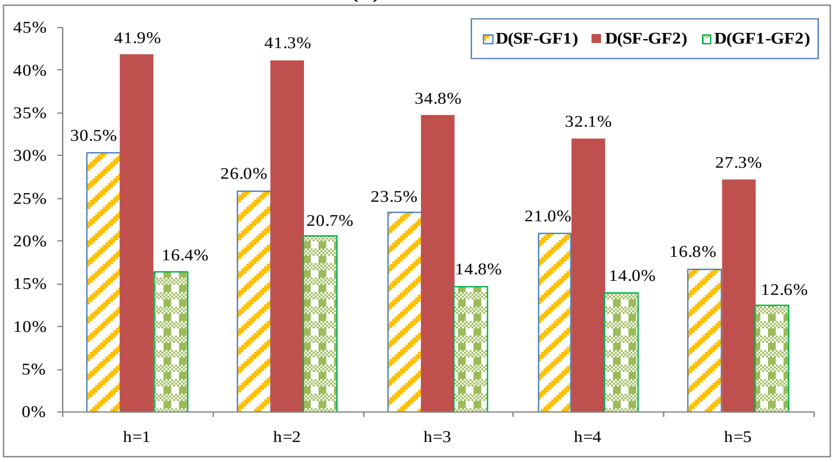
As the forecasting horizon extends, uncertainty grows and forecasting is expected to become increasing more difficult as the range of alternatives becomes large and cumbersome. Table 5.13 shows that the accuracy of the group forecasts decreased over time from 2011Q2 to 2012Q2. During the quarter that the experts were asked to make forecasts, they had more information to make an evaluation of the present forecasts than they had to evaluate future forecasts, and this would have enabled them to produce more accurate short-term forecasts than long-term forecasts.

However, the group forecasts in Round 2 for h = 2 to 3 were more accurate than the current quarter (h = 1) as the MAPE of the two- (4.35%) and three-step-ahead (5.7%) were less than the current-quarter forecasts (6.93%). Several reasons may help to explain this finding. First, the baseline forecasts of h = 2 to 3 were more accurate than the current-quarter forecasts. This accuracy difference was likely to be reflected in the revised forecasts. Second, the occurrence of a disaster event like the Japan earthquake in March 2011 added to the difficulty of accurately predicting the demand from the Japan market and its competing destinations.

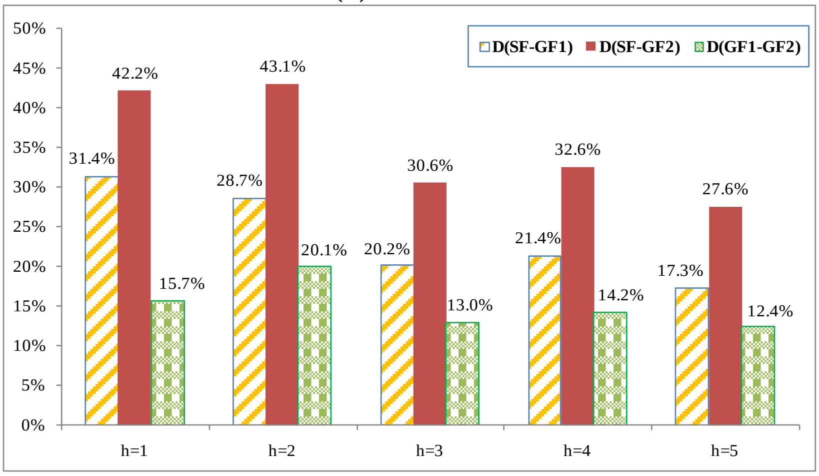
The degree of improvement over the accuracy criterion was measured by the percentage reductions in the accuracy measures of the revised forecasts over the initial forecasts. Figure 5.26 shows that there was an overall decreasing trend of improvement in the accuracy measures over the period 2011Q2 to 2012Q2. Furthermore, greater accuracy improvements were found after the second round adjustment than after the initial round; for example, the accuracy improvement of the Round 1 group forecasts over the initial statistical forecasts in RMSPE decreased from 30.5% (h = 1) to 16.8% (h = 5). Even larger improvements, from 41.9% (h = 1) to 27.3% (h = 5), were observed after the experts' adjustment in the second round.



(a) MAPE



(b) RMSPE



(c) Theil's U statistic

Figure 5.26 Improvement in accuracy by forecasting horizon

Note: Accuracy improvement is measured by the percentage reduction (*D*) between the accuracy measures of two forecasts; for example, to calculate the percentage reduction of MAPE: $D = (MAPE_{SF} - MAPE_{GF1}) / MAPE_{SF}$.

The above analysis utilized the information concerning percentage error reductions, while the following analysis provides information regarding error reductions in APE. Table 5.29 provides the distribution information of APE for three sets of forecasts (SF, GF1, and GF2) by quarter. A similar decreasing trend in accuracy improvement over time can be observed in this table, and this supports hypothesis *H5d* that improvement in accuracy resulting from experts' adjustment decreases over time.

It is worth noting that the evaluation results were somewhat distorted by comparing the performance of quarterly forecasts, which may dilute the pattern of decreasing accuracy over time. Because annual predictions may capture the overall trend that is expected to prevail in the next year but fail to correctly project the quarterly movements (Joutz & Stekler, 2000), evaluations should be conducted based on annual data where the errors made in predicting quarter-to-quarter changes might cancel themselves out.

Table 5.29 Descriptive statistics of APEs by forecasting horizon

		Panel A				Panel B	
Quarter		Min (%)	Max (%)	Mean (%)	Std. Deviation	Mean Difference	Improvement (%)
2011Q2	APE_{SF}	1.93	28.82	10.64	0.11	GF1-SF	3.20
	APE_{GF1}	0.42	18.47	7.44	0.08	GF2-SF	3.71
	APE_{GF2}	2.09	14.27	6.93	0.05	GF2-GF1	0.51
2011Q3	APE_{SF}	0.38	26.88	7.01	0.10	GF1-SF	0.67
	APE_{GF1}	1.34	18.20	6.34	0.06	GF2-SF	2.66
	APE_{GF2}	0.08	14.79	4.35	0.06	GF2-GF1	1.99
2011Q4	APE_{SF}	1.92	20.00	6.53	0.07	GF1-SF	0.32
	APE_{GF1}	1.16	11.10	6.21	0.03	GF2-SF	0.83
	APE_{GF2}	3.10	7.56	5.70	0.02	GF2-GF1	0.51
2012Q1	APE_{SF}	0.16	26.87	8.77	0.09	GF1-SF	0.79
	APE_{GF1}	0.99	17.59	7.98	0.06	GF2-SF	1.71
	APE_{GF2}	2.07	15.39	7.06	0.05	GF2-GF1	0.92
2012Q2	APE_{SF}	0.76	38.30	9.98	0.14	GF1-SF	0.25
	APE_{GF1}	0.75	29.85	9.73	0.11	GF2-SF	1.66
	APE_{GF2}	1.12	26.46	8.32	0.09	GF2-GF1	1.41

Note: Improvement in Panel B represents the mean difference in the accuracy of the adjusted forecasts over the initial forecasts.

(6) Hypotheses about adjustment behaviour

H6a: The forecasts selected for adjustment are those most in need of adjustment.

Although the forecasts for all of the six markets were revised by the experts, the magnitude of their revisions varied considerably among the markets. Table 5.30 gives some descriptive statistics on the relative size of the forecast revisions in arrival units. It can be seen that for both rounds, the positive adjustments were very much larger than the negative adjustments. The Mainland market was found to be adjusted most frequently, followed by Australia and the USA. The UK market was the least revised in terms of both positive and negative adjustments. In addition, the frequency of positive adjustments appeared to be lower with iteration, while negative adjustments did not change too much over rounds.

Table 5.30 Direction of relative adjustments to the baseline forecasts

Market	**R1**				**R2**			
	Negative (a)	No	Positive (b)	Total (=a+b)	Negative (a)	No	Positive (b)	Total (=a+b)
China	34	70	238	272	27	151	145	172
Australia	60	92	190	250	56	176	91	147
USA	70	125	147	217	64	217	42	106
Japan	87	129	126	213	80	221	22	102
Taiwan	48	156	138	186	43	229	51	94
UK	53	193	96	149	49	263	11	60
Total	352	765	935	2052	319	1257	362	1938

Note: Negative: negative adjustments made to the baseline forecasts, No: no adjustments made to the baseline forecasts, Positive: positive adjustments made to the baseline forecasts.

The process of graphical adjustment is inherently appropriate in the sense that adjustments are most helpful when they are most needed. This conclusion relies on the concept of "excess error", which is defined as "the difference in error of two alternative methods, one being the simplest Naive forecast" for real data (Willemain, 1991, p. 154). If a forecaster is presented with a graph of the data and forecasts, his/her adjustment will be a reflection of the combined effects of substantive expertise and, perhaps unconsciously, graphical adjustment (Willemain, 1991). Willemain concluded that studies of judgmental forecasting adjustment could reach different conclusions depending on whether the excess errors are positive or negative for the majority of their data series.

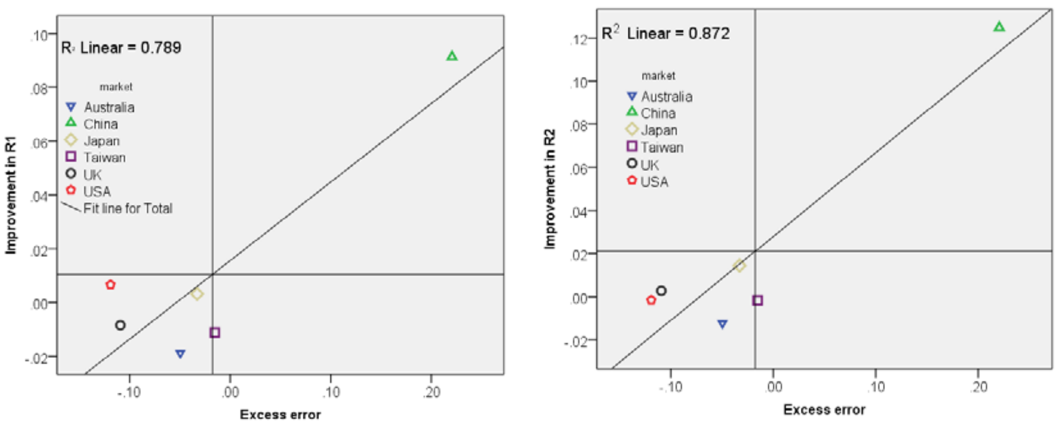
The concept of excess error was employed to test *H6a*. Table 5.31 and Figure 5.27 show the relationship between accuracy improvement and excess error. For both rounds, it was found that improvement in accuracy was closely associated with excess error. The Pearson correlation coefficient between improvement in accuracy and excess error was 0.89 for forecasts in Round 1, and this positive association was stronger for forecasts in Round 2 as the coefficient increased to 0.93. The p values for the above two correlation tests were far below 0.01, indicating that the positive relationship between accuracy improvement and excess errors was statistically significant at the 1% level.

The findings from the above correlation analysis suggested that the arrivals series with greater accuracy improvement generally had a greater excess error and that this relationship became more evident after iteration, which supports the view expressed in *H6a*. This finding is in accord with that of Willemain (1991), who concluded that excess error was the only variable that discriminated between helpful and harmful adjustments. Willemain also provided one reason to explain the possible conflicting results obtained from different studies of judgment in forecasting, namely that different studies might unwittingly operate with different values of excess error. It is noted that the results for these tests in this study should be interpreted with some caution due to the limited arrival series (only six) employed to conduct the analysis.

Table 5.30 also shows that the experts made adjustments to the statistical forecasts for all six source markets, even those markets with high accuracy such as Australia and the UK. This is consistent with the findings of Lim and O'Connor (1995), who found that even when an impossibly good forecast is provided, forecasters still make small adjustments to it even if such adjustments have no sound basis, and therefore they diminish the forecast accuracy.

Table 5.31 Improvement and excess error for group forecasts by market

Market	Excess error (%)	Improvement (R1, %)	Improvement (R2, %)
Australia	-4.96	-1.88	-1.23
China	22.06	9.13	12.48
Japan	-3.29	0.32	1.45
Taiwan	-1.53	-1.11	-0.16
UK	-10.93	-0.84	0.28
USA	-11.91	0.66	-0.15
Pearson Correlation		0.89	0.93
***p*. (1-tailed)**		0.009	0.003



(a) R1 (b) R2

Figure 5.27 Improvement and excess error for adjusted forecasts over rounds

H6b: The size of forecast adjustment will be associated with the direction of forecast adjustment.

The previous hypothesis examined the relationship between accuracy improvement and excess error in order to investigate the validity of experts' judgmental behaviour (whether those forecasts selected for adjustment are those most in need for adjustment). This section aims to explore the relationship between the size of forecast adjustment and its direction.

With iteration, the number of "no adjustment" increased from 193 in Round 1 to 320 in Round 2, while the revision activities of "positive adjustment" and "negative adjustment" tended to shrink over rounds. To obtain a deep insight into the relationship between the size and the direction of adjustments, the size of the adjustments was recoded based on quartiles. In each round, the majority of the larger adjustments were positive adjustments while the smaller adjustments appeared to be negative adjustments (see Table 5.32). From the cross tabulation alone, it is impossible to decide whether this difference was real or due to chance variation. The Chi-square test was thus carried out to examine the existence of such a difference. The two-sided asymptotic significance of the Chi-square statistic was less than 0.01 (see Table 5.32) in both two rounds, and so it is safe to conclude that the difference was statistically significant, which implies that there is an association between the size of adjustments and the direction of adjustments.

The Chi-square test is useful for determining whether or not a relationship exists, but it cannot help to quantify the strength of the relationship; therefore, symmetric measures, including Phi, contingency coefficient, and Cramer's V, were also used. The results in Table 5.32 show that all three measures were significant at the 1% level irrespective of the Delphi round being examined, which indicates that the strength of the relationship was significant. Moreover, the large values of all three measures (all above 0.7) suggest the existence of a strong relationship. The results from the Chi-square test and the symmetric measures support hypothesis *H6b*.

Table 5.32 Breakdown of adjustments in the size and direction over rounds

R1				
Size of adjustment	**Direction of adjustment**			Total
	Negative adjustment	No adjustment	Positive adjustment	
Below Quartile 25%	0	193	0	193
Quartile 25%-50%	40	0	37	77
Quartile 50%-75%	41	0	93	134
Quartile 75% and above	13	0	123	136
Total	94	193	253	540
Pearson Chi-Square	611.721**			
Symmetric Measures				
Phi	1.064**	Cramer's V	0.753**	
Contingency Coefficient	0.729**			

R2				
Size of adjustment	**Direction of adjustment**			Total
	Negative adjustment	No adjustment	Positive adjustment	
Below Quartile 25%	0	320	0	320
Quartile 25%-75%	36	0	26	62
Quartile 75% and above	51	0	77	128
Total	87	320	103	510
Pearson Chi-Square	524.995**			
Symmetric Measures				
Phi	1.015**	Cramer's V	0.717**	
Contingency Coefficient	0.712**			

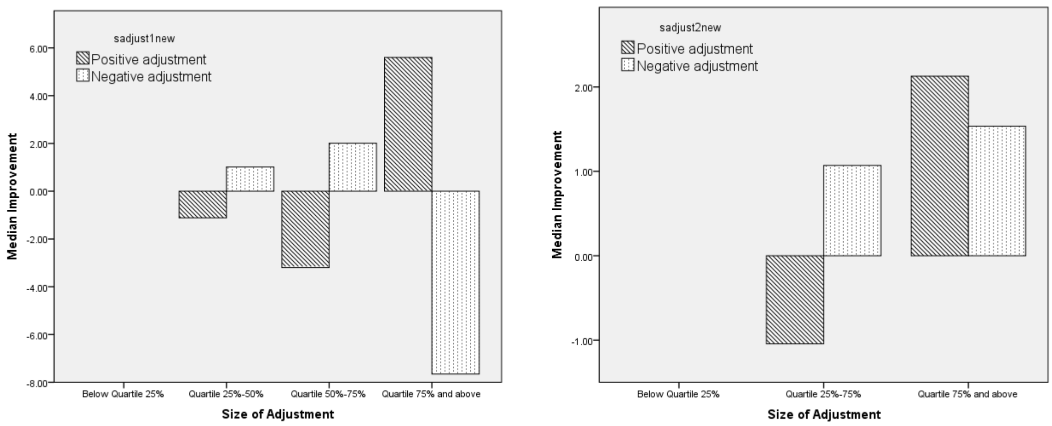
Note: ** and * indicate significance at the 1% and 5% level, respectively.

H6c: When adjustments are made, the size of the judgments is positively associated with an improvement in accuracy.

Linstone and Turoff (2002) showed that different people have their own preferences for future outcomes: some Delphi participants are inherently optimistic whereas others are pessimistic. These tendencies (either being overly pessimistic or optimistic) are complicated by individual characteristics and are likely to affect participants' final forecasts. Armor and Taylor (2002) illustrated the outcomes of making pessimistic predictions and optimistic predictions: Overly pessimistic predictions may be demoralizing if they are believed, and if they are fulfilled, the outcomes that are obtained may not be very satisfactory; however, overly optimistic predictions may confer benefits simply by symbolizing a desired image of success.

Given that predictions are often inaccurate, at least to some degree, it is possible that people may derive benefits from shifting the range of their predictions to the positive even if this means introducing an overall higher rate of forecast errors.

This hypothesis test aimed to examine whether the size of judgmental adjustments was related to accuracy improvement. To test *H6c*, the IMP measure was first calculated according to Equation (3.10). Figure 5.28 shows the median improvement in APE for different sizes of relative adjustments. It seems that there is no clear pattern from which to determine the relationship between the size of adjustment and accuracy improvement.



(a) R1 (b) R2

Figure 5.28 Improvement in accuracy by adjustment size

In Round 1, the middle 50 per cent of the negative adjustments − the size of adjustment fell between quartiles 25 to 75 per cent − on average led to higher forecast accuracy. By contrast, positive adjustments, on average, reduced accuracy, although they appeared to increase accuracy for the largest 25 per cent of adjustments.

In Round 2, the negative adjustments seemed to be more beneficial than the positive adjustments as they increased accuracy for half of the adjustments. These findings conflict somewhat with those of Mathews and Diamantopoulos (1992).

Based on a large warehouse operation where forecasts were routinely adjusted, Mathews and Diamantopoulos found that marginal adjustments were often more beneficial than large adjustments but were also biased.

In short, hypothesis *H6c* cannot be confidently accepted based on the above findings. It would thus be worthwhile investigating the relationship with a longer span of data.

The findings from testing hypotheses *H6a*, *H6b*, and *H6c* indicate a need to record the history of user interaction with the forecasting system. Armstrong and Collopy (1998) strongly suggested that forecasters should keep accurate records of the magnitude of the adjustments they made, the process they used to make the adjustments, and the reasons for making them. This documentation process provides a number of benefits. First, this feedback process can have a powerful effect on the accuracy of forecasts. Sanders and Ritzman (2001) stated that the documenting process would help experts to see the effects of specific types of judgmental adjustments, learn from their past forecasts and improve their future forecasts, and meantime develop their forecasting expertise. Studies have shown that good feedback can improve forecasters' learning and improve their performance of most estimation tasks (O'Connor, 1989). Second, reviewing these records can also serve to discourage politically motivated biases, which can be intentional in nature. Third, the documenting process should help forecasters to use contextual knowledge in a more effective manner as they may devote more thought to the judgmental adjustment when they are documenting the process (Sanders & Ritzman, 2001).

5.6 In-depth Interviews: Findings and Implications

5.6.1 Introduction

The purposes of the in-depth interviews in this study were to explore the underlying assumptions for the experts' Delphi forecasting, investigate the causes of bias and inefficiency caused by judgmental adjustments, gain insights into the experts' judgmental forecasting behaviour, and seek recommendations for future improvements to the HKTDFS from tourism researchers and practitioners.

This section summarizes the findings from the in-depth interviews by tabulating and categorizing the participants' comments and views. It is divided into seven main categories (see Figure 5.29) or 13 subcategories.

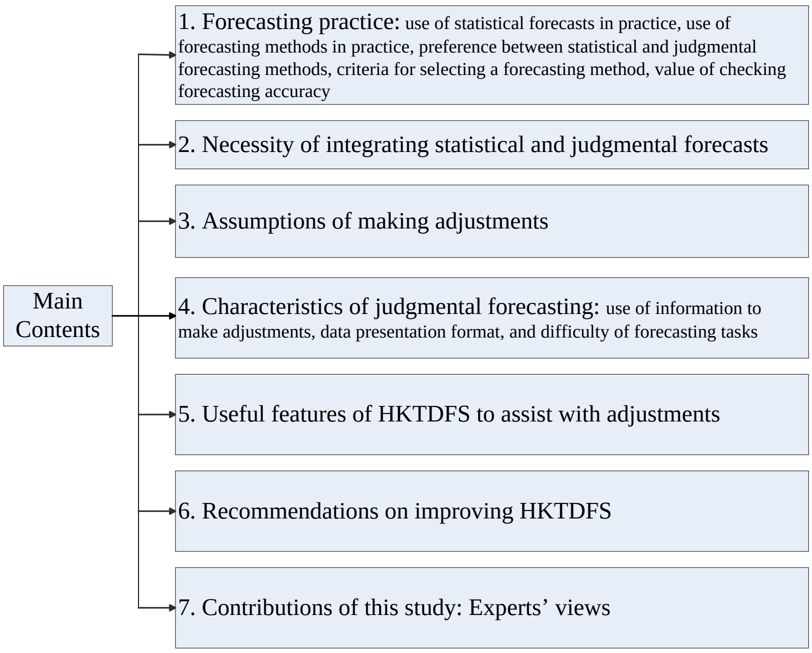


Figure 5.29 Analytical structure of in-depth interview findings

The first category summarizes experts' comments related to forecasting practice in the respondents' organizations, such as the use of forecasting methods in practice, the forecasting method preferences of individual tourism forecasters, the criteria used in selecting a forecasting method, and perceptions regarding the value of evaluating forecast accuracy. The second category analyses the comments related to the necessity of integrating statistical and judgmental forecasts in tourism. The third category explores the experts' comments on the assumptions of judgmental adjustments for Delphi surveys. The fourth category investigates the characteristics of judgmental forecasting, including the types of information used to assist expert judgments, the format of data presentation, and the difficulty of conducting forecasting tasks. The fifth category examines the reasons for accuracy improvement by summarizing the experts' views regarding the useful features of the HKTDFS that help with judgmental adjustments. The sixth category summarizes the experts' recommendations on how to improve the functional ability and forecasting performance of the HKTDFS. The last category explains the contributions of the current study from the experts' perspective. To obtain insightful information, an analysis is conducted on the basis of the 13 subcategories in Section 5.6.2.

5.6.2 Key findings

(1) Use of statistical forecasts in practice

At the start of each interview, respondents were asked to indicate if they used statistical forecasts in their forecasting practice. Eight of the nine academic researchers had experience of using statistical methods to predict demand; one of them only had teaching experience. All five industry experts indicated that their organizations used statistical methods to predict tourism demand.

Furthermore, the industry experts indicated that they did not rely on a single source of forecasts but rather used forecasts from external organizations as well. For example, the four most frequently mentioned sources used to obtain forecasts were UNWTO, IMF, Pacific Asia Travel Association (PATA), and the HKTDFS. Two experts mentioned that they sometimes also used commercial data services, for example, the Economist Intelligence Unit (EIU). In addition, as stated by one expert, when they used forecasts from external organizations, they seldom challenged the forecasting methods; however, they always asked for such information as the key variables and factors included in the models.

All of the industry experts indicated that in practice they usually took statistical forecasts as references or as benchmarks for their strategic planning. For example, one tourism officer indicated that, in his organization, he and his colleagues adopted statistical forecasts for reference only and made their own forecasts based on a set of assumptions according to market information. Another tourism officer shared a similar view, but he placed his trust and emphasis on long-term forecasting. This expert believed that it was too risky to rely only on statistical forecasts, particularly in the long term because there are so many different factors impacting long-term forecasting.

As described earlier, the respondents revealed that in their work, they use statistical forecasts as the baseline for making their adjustments. At a subsequent stage, they incorporate opinions and comments from other experts to reach their final predictions. In other words, their final forecasts are a combination of business insights and statistical methods. According to four of the five industry experts, they use a broadly similar process to make their final predictions: At the start of each forecasting period, statistical forecasts are produced. Then, a forecasting meeting is held, generally involving forecasting, marketing, production, and sales personnel; at this meeting, statistical forecasts are examined in the light of marketing and other relevant information; and final forecasts are agreed on. Examples of specific comments are provided below (see Box 5.1).

Box 5.1

- *Sometimes, we do not simply base our forecasts on statistical methods. We include how the senior management feels [the] business is going, particularly regarding sales because they have the market insights and know what the business is heading into. [Industry]*
- *We produce our own statistical forecasts and then our management adds their insights and we adjust the forecasts accordingly. [Industry]*
- *Particularly with regard to sales, when sales people talk to travel agents or wholesalers, they will know that their business is slowing down; particularly in the case of airlines, they have knowledge of advance bookings, and so they will know what the business will look like in about three or six months. [Industry]*
- *I rely very much on what I hear from hotels, restaurants, and airlines. I rely solely on this kind of data. I listen to those people with open ears. I may not believe 100 per cent of what they say. [Industry]*
- *We see if these forecasts make sense, and the forecasters will need to provide rationales for their forecasts. [Industry]*
- *Furthermore, I must say that a good forecaster or a good economist should never believe any model results without recalling or proving that the model is practical. [Academic]*
- *I would say that sometimes the result of a forecast has to align with our business sense as well. We have to admit that when it comes to the business world, science is science and reality is reality: [the] business world is reality. Therefore, we have to combine these two. Even though we are talking about arrivals from China, all we thought is that China is going to grow at 3 per cent according to our model. Even if this is accurate, if the whole industry or our whole senior management group does not believe that it will only be 3 per cent, but rather believe that it will be 5 or 7 per cent, we will adjust the forecast. So sometimes the model has to go with the business perception of the world. [Industry]*

(2) Use of forecasting methods in practice

As revealed earlier, the experts generally took statistical forecasts as the references for making their adjustments. This section provides information on the use of forecasting methods among 14 experts. Table 5.33 shows that the most widely used forecasting method was regression as 11 out of 14 experts mentioned that they used regression methods to produce their forecasts.

It seems that academic experts' and industry experts' usage of forecasting methods is quite different. Generally, compared to the industry experts, the respondents from academia seemed to favour more complicated/sophisticated forecasting methods (see Table 5.33). The forecasting methods used by the academic interviewees included regression methods (9), econometric models (6), time series models (6), artificial intelligence methods (2), data mining techniques (1), and some benchmark models (3) such as Naive models, moving averages, and exponential smoothing methods. In addition, one academic expert briefly summarized the methodological development trend in tourism demand forecasting since the 1990s− “compared with the published studies prior to 2000, forecasting methodologies have become more diverse in the new millennium.” He also argued that the choice of a forecasting method also depended on the specific problems, the purpose of forecasting, and the data availability.

Table 5.33 Forecasting methods used by tourism forecasters

ID	Category	Forecasting methods
1	Industry	Analysis on historical trends and adjustments from senior management
2	Industry	Judgmental forecasting: gut feelings based on historical data (e.g. arrival trend, business trend, travel trend, etc.), economic data (e.g. GDP growth, income level), business environment, availability of flights [from source markets to Hong Kong], transportation system in Hong Kong, and data in other travel-related business (travel trades including travel agents, tour operators, wholesalers, airlines, etc.)
3	Industry	Regression models (e.g. linear regression and logistic regression) and adjustments from senior management
4	Industry	Regression methods and adjustment from senior management
5	Industry	Analysis of historical data (e.g. trends, growth rates) and adjustments from senior management
6	Academic	Time series models (ARIMA, seasonal ARIMA, Naive models), data mining techniques (rough sets), and simple econometric models (e.g. error correction model)
7	Academic	Econometric method, e.g. TVP, ARDL, VAR
8	Academic	Regression analysis
9	Academic	Regression analysis
10	Academic	Time series models, causal econometric models, expert opinions
11	Academic	Traditional methods (e.g. Naive 1, Naive 2, moving average, exponential smoothing, seasonal ARIMA, and multiple regression) and artificial intelligence methods (e.g. artificial neural network, rough sets, etc.).
12	Academic	Econometric methods (e.g. ARDL, ECM, TVP), time series models (e.g. ARIMA), Naive models and exponential smoothing methods for comparison purposes.
13	Academic	Econometric methods (e.g. ARDL, ECM), time series models (e.g. ARIMA)
14	Academic	Simple time series models and econometric methods

Unlike the academic respondents, the forecasting methods reported by the industry experts included both quantitative forecasting techniques, such as regression methods (2) and the analysis of the trends and growth rates in historical data (2), and judgmental forecasting (1). A typical observation was that judgment is commonly used in practice among tourism forecasters in the industry to complement their use of relatively easy forecasting methods. For example, one industry expert admitted that his organization's methodology was relatively simple and easy to understand and even arbitrary, but he stated that his organization included judgments to take into consideration short-term variations in the market as well as longer-term impacts.

An operational definition of simplicity is as follows: "the approach can be explained to a manager so clearly that he could then explain it to others" (Armstrong, 2001e, p. 375). Simplicity in an econometric model means a small number of causal variables and a functional form that is linear in its parameters. Simpler forecasting methods also help decision-makers to understand and implement their forecasting tasks, which could possibly reduce the likelihood of mistakes and errors, and are less expensive (Armstrong, 2001e).

As discussed previously, the industry experts usually took statistical forecasts as a starting point to assist with their final forecasting decisions. They would try to incorporate as much market-relevant information into their final forecasts as possible. This practice is highly difficult to achieve using statistical forecasting methods. For example, one industry expert who had more than 20 years' forecasting experience reported that he would not take a statistical forecast as "gospel" but as a reference value and said that in making his final forecasts, he mostly relied on his "gut feelings" to combine all useful market information. One academic expert stated that the difference between academic researchers and industry practitioners is that the methods used for academic research can be “extremely complicated” whereas the methods designed for industry practitioners should be “easy to use”.

In addition, the industry respondents stated in the interviews that they did not rely on a single forecasting method all the time. For example, one interviewee stated that he used regression models in most cases and sometimes simply used the growth rates in recent years as a reference to produce forecasts. Another interviewee commented that the selection of a forecasting method depended on the actual needs. She further illustrated how her company selected statistical forecasting methods by using the statistical criteria (e.g. MAPE) embedded in a commercial system (SAS). The following quote illustrates how this expert made statistical forecasts in her company.

> *Then they (the computer program) will list out all these different methods and we will choose which one it makes more sense. ... So instead of dealing ourselves, the computer has actually helped to do the models. I think for **SAS**, it already has like **50 different methods** and they just run it and show us the **MAPE** or whatever the criteria it sets up.*

(3) Preference between statistical and judgmental forecasting methods

The previous section revealed that the industry experts used both statistical forecasting methods and senior judgments to adjust forecasts, while the academic experts used more advanced and sophisticated statistical forecasting methods but seldom incorporated human judgment in their final forecasts. This section provides further details about the experts’ preferences in terms of selection of forecasting methods and explores the reasons behind their selection.

Eight of the nine academic researchers interviewed indicated a preference for using statistical forecasting methods rather than expert judgments. One explanation they gave for this preference was that they were more familiar and comfortable with quantitative forecasting methods. For example, one expert stated that he preferred statistical methods because they were (a) the ones he was most familiar with and (b) “well-grounded in the sense that there are a set of models and assumptions [upon which] to base forecasts”. Two other experts explained that their preference for quantitative forecasting methods was simply based on the fact that they had not used any judgmental forecasting methods before. Based on the above observations, it seems that the reason why the academic experts preferred quantitative forecasting models was because they were trained in their use. This phenomenon has been observed in the existing literature. For example, Armstrong (2001e) indicated that there is a common presumption that researchers who are skilled at a technique will force their technique on the problem at hand. However, a method selected on the basis of convenience may be difficult to understand (Armstrong, 2001e). Furthermore, Armstrong (2001e) also cautioned against the selection of a forecasting method on the basis of convenience as this would probably result in serious forecast errors in situations that involve large changes.

As revealed previously, the academic experts usually adopted more advanced and sophisticated forecasting methods whereas the industry experts tended to choose much simpler methods. One industry expert stated straightforwardly that “if we want to go from point A to B, we want to get there as simply as possible”; if a forecasting method is difficult to use, they would not use it because it would create troubles rather than benefits.

Despite the different preferences among the experts with regard to forecasting methods, they generally agreed that forecasting methods should be appropriately selected under different conditions. For example, one academic expert indicated that he used both statistical and judgmental forecasting methods depending on the purpose of the forecasts or the situation: "When I am doing more scientific papers, I use statistical models. When I am doing more forecasting with political impacts, I use the opinions of experts." Another academic expert regarded statistical forecasts and expert judgments as complementary rather than as substitutes. He thought that expert judgments should be applied on the top of statistical forecasts because he believed that statistical forecasts provide a solid and reliable foundation and that judgmental adjustment was necessary.

(4) Criteria for selecting a forecasting method

The selection of criteria represents a critical step in the evaluation of forecasting methods. It is thus of interest to examine what criteria tourism forecasters use when selecting forecasting methods. Table 5.34 summarizes the selection criteria that were mentioned by the 14 experts interviewed. The results are similar to those from the previous studies reported in Armstrong (2001e).

Table 5.34 Criteria for selecting a forecasting method

Criteria (Frequency)	
Accuracy (13) • So the **accuracy** is the first concern. • Because you will make important decisions based on the forecasts, so it has to be **accurate** in order to ensure your decisions are relevant and appropriate. • ... the **first** thing is the possibility of generating a **good** forecast. • The **first** thing is the **accuracy** of the forecasts. • For forecasting, the most **important** thing is **accuracy**. • **Accuracy** is a must.	**Others:** • Cost/time (7) • Ease of use/implementation (5) • Ease of interpretation (4) • Robustness (3) • Adaptive to new conditions (2) • Purpose of study/project (2) • Data availability (1) • Researchers' skills (1) • Capture turning points (1) • Ability to evaluate policies (1) • Transparency of the method (1) • Possibility to duplicate forecasts (1) • Stability (i.e. on more occasions the same method can show good forecasting performance.) (1) • Incorporates judgmental input (1) • Speed with which the forecasts are available (1)

The interviews revealed that the academic researchers and industry practitioners had similar views on the importance of various criteria. Among the 14 respondents, 13 (5 industry experts and 8 academic experts) indicated that accuracy was the most important criterion in selecting a forecasting technique (see Table 5.34). Moreover, these 13 experts also agreed that one principal objective of the forecasting process is to produce accurate forecasts.

Although accuracy is usually the primary concern, other criteria should also be considered. Table 5.34 shows that the second most frequently mentioned criterion was the timeliness and cost of making forecasts. The amount of time available should be considered in relation to the total forecasting effort; when forecasts for a large number of items are required, it can mean that a considerable time commitment is needed for forecasting. Thus, although the other criteria may indicate that the application of a sophisticated method would be worthwhile in a given situation, the amount of time available may lead to the application of a less sophisticated method simply to save time. As mentioned by several respondents, another important factor influencing their selection of a forecasting method is the limited amount of time in which to make a decision, and thus they would tend to select a simpler and easier method.

Among the other criteria that the experts mentioned were ease of use (implementation), ease of interpretation, robustness, adaptive to new conditions, purpose of forecasting, data availability, incorporation of judgmental inputs, and ability to evaluate policies (see Table 5.34).

Ease of use/implementation relates to how well tourism forecasters can understand the method, how valuable its results are to them personally, and how easily they are able to implement the method. Complex and highly mathematical methods are generally less appealing than simpler techniques they can understand without a tremendous amount of training. Two criteria − ease of use/implementation and ease of interpretation− were regarded as being nearly as important as accuracy by the industry experts. This is consistent with the finding from Yokum and Armstrong's (1995) study that forecasting experts, especially practitioners, regard ease of use as almost as important as accuracy. In addition, Armstrong (2001e) concluded that ease of interpretation and ease of use are considered to be equally as important as accuracy. Actually, a forecasting method with high accuracy will be of little use if it is rejected or misused by practitioners.

In addition to the criteria listed above, organizational pressure is another criterion that two tourism officers considered when selecting a forecasting method. These two government officers shared the same view that official tourism forecasts may be subject to organizational pressure to incorporate a forecast bias. They thought that official tourism forecasts tended to be too conservative (or underestimated) due to political as well as the budgetary concerns. Lawrence, O'Connor, and Edmundson (2000) found that political and organizational structures and incentives encourage forecasters to take account of the consequences of forecast errors. The high uncertainty of the tourism industry is another possible reason uncovered by the experts. Caiden and Wildavsky (1974) concluded that the higher the uncertainty, the higher the conservatism.

(5) Value of checking forecast accuracy

When asked about the criteria used in selecting a forecasting method, the experts agreed that accuracy was the most important criterion in assessing the forecasting process. In this section, the experts' comments regarding the value of checking forecast accuracy are examined.

The experts generally agreed that it is necessary to check forecast accuracy and that this should be encouraged in practice in the evaluation process. For example, one expert explained that forecast accuracy is regarded as the justification of the predictions they make when presenting them to senior managers. Another expert stated that two reasons for forecast evaluation are to "learn something to improve the model" and to "learn more how the reality behaving" (i.e. how the forecasts compare to the actual values).

Typical comments made by the industry interviewees expressed a lack of good procedures for evaluating forecasting methods in practice and difficulty in keeping tracking records for the different forecasting methods used. All five industry experts stated that they conducted *ex ante* forecasting. Two of them indicated that they used MAPE and RMSE to check the accuracy of hold-out data which was then used to select the final forecasting model. The other three experts stated that to compare forecast accuracy, they directly compared actual figures with their forecasts (but only for recent quarters or years) and provided reasons if there was a big discrepancy. The three experts also mentioned that their organizations kept updating their forecasts to incorporate the latest information in order to improve forecast accuracy. One expert expressed that tourism demand is highly volatile and that therefore it is critical to closely monitor the changes in the market environment and incorporate the latest information into forecasts in a timely manner.

In addition, a tracking record of forecasting performance (accuracy) is likely to help forecasters enhance their self-confidence. For example, one expert explained: "if I am close to be[ing] accurate over the last 3 years, then I build confidence in myself to make the right decision."

Four of the nine interviewees from academia indicated that they conducted *ex post* forecasting with in-sample forecasting evaluations. Two of them reported that they conducted both *ex ante* and *ex post* forecasting. Another emphasized that *ex ante* forecasts were much better than he expected. He also revealed that advanced econometric models performed equally well for both *ex ante* and *ex post* forecasts. In his studies, he used MAPE to evaluate forecast accuracy.

(6) Necessity of integrating statistical and judgmental forecasts

The study by Sanders and Titzman (2001) suggested that it is sensible to integrate judgmental and statistical forecasting methods as each have their strengths and weaknesses. This section examines the experts' opinions regarding the necessity of integrating statistical and judgmental forecasts in the tourism forecasting context. Three types of information were collected—namely respondents' perceptions regarding the need to integrate statistical and judgmental forecasting methods, the benefits of integrating these two forecasting methods, and reasons for integration, and these are analysed in this section.

All of the respondents agreed that given the relative strengths and weaknesses of judgmental and statistical forecasting methods, it is necessary to integrate these two types of forecasts in tourism forecasts. A few interviewees indicated that such integration is very beneficial for the tourism industry. They also illustrated that such integration would be particularly useful in Hong Kong and many other Asian destinations where tourism demand is full of dynamics and uncertainties. Leitner and Leopold-Wildburger (2011) suggested that when faced with such scenarios, relying solely on the output of a statistical forecasting model would be suboptimal and judgmental modification by decision makers should also be required.

The majority of the experts interviewed addressed the benefits of combining experts' knowledge with statistical forecasts. Table 5.35 summarizes the two main benefits mentioned in the interviews − improve d forecast accuracy and the integration of experts' knowledge into the forecasting process. For the respondents, the primary benefit was improved forecast accuracy. Typically, they commented that the integration of the two methods would lead to more reliable and accurate forecasts and would reduce the forecasting risk.

The second benefit is the integration of experts' contextual knowledge regarding the latest changes in the market environment and any factors impacting the variables being forecast into the forecasting process. According to the experts' views/comments, the integration of human judgment into statistical forecasts can be rationalized in two ways. The first rationale is that people can incorporate up-to-date knowledge on changes and events that are occurring in the environments that can affect the variable being forecast and/or past experience into the forecasts. Practitioners often have current (i.e. non-time series) information that cannot be captured by statistical models but may possibly lead to improvements in forecast accuracy (Edmundson, Lawrence, & O'Connor, 1988). The second rationale for the incorporation of human judgment into statistical forecasts is that people might be better able to detect changing patterns in a time series than statistical forecasting methods. One expert believed that the failure of statistical forecasting methods to capture turning points made it appealing and valuable to involve experts' judgments in the forecasting process.

Table 5.35 Summary of experts' opinions on integration

Item	Examples of Comments
Improve forecast accuracy	• I do believe that combining the two or including expert forecasts based on statistical forecasting will lead to more reliable and more accurate forecasting results. [Academic] • I do believe [that] the combination will lead to very reliable and accurate forecasts. [Academic] • I believe that using a combination of different methods may give us a better forecast. [Academic] • So if we combine statistical forecasts with expert opinions or expert adjustments, it may improve forecast accuracy. [Academic] • So it is necessary to combine these two [statistical and judgmental methods] to generate more accurate forecasts. [Academic] • I think [that] probably the combination results may be better or more accurate than the worst single one. [Academic] • I think it is necessary to combine because if the combination can outperform the worst single one, then the practice of forecasting will be less risky. [Academic] • I think that combining what people what are actually seeing with what it is happening raises the accuracy. [Industry] • I think that to increase accuracy or to get better accuracy, the two should be combined. [Industry] • We just think that combination is more accurate… [Industry] • You need to figure out what formula, which method, and what model can produce accurate and useful forecasts. Accurate and useful, those are the two things. [Industry] • I think it is better to integrate these two methods (statistical and judgmental methods). There are some factors which cannot be captured by statistical models. Integration can make two forecasting methods complement each other, given the relative advantages of each method. [Industry]
Integration of experts' knowledge into the forecasting process	**(i) Integrate up-to-date information** • ... the market is growing and is full of dynamics. In those cases, it is very necessary to bring in experts' opinions, based on their knowledge about the future, about current tourism development situations in those countries and their beliefs about future growth. Then these beliefs should be brought into the forecasting process to adjust the statistical forecasting results. [Academic] • ... regular forecasting methods are not be able to take account of new conditions, whereas expert judgments can take them into consideration. [Academic] • For example, Mainland China suddenly introduced new policies for outbound tourism. If they introduced the policy today, and if this policy is effective immediately, no forecasting method could capture this impact, but human experts can. [Academic]

Table 5.35 Summary of experts' opinions on integration (Continued)

Item	Examples of Comments
	• …, we need human judgment to adjust the number [of arrivals] down. [Academic] • … take the recent economic crisis for example. If you simply generate forecasts based on the econometric models, it may not take the current economic situation into account because when you generate those forecasts using historical data, it may not reflect the future trend. But if you ask the expert, he or she may have some personal insights into the current or future economic environment, and they will take this into account. So their predictions can somehow take this kind of information into consideration, and thus their insights will be valuable when you want to generate more accurate forecasts. [Academic]
	(ii) Detect changing patterns in the time series • But if there is a turning point in terms of the growth rates, which has not been reflected in the historical data, experts may be able to capture it. [Academic] • … I think past data cannot catch the trend, nor is not able to tell us the future trend. In that case, expert inputs are very important. [Academic] • The art of forecasting is to identify trends for downturn when numbers are up and trends for upturn when numbers are down. [Industry]
Statistical and judgmental forecasting are substitutes.	• We do not predict that any unusual things will happen, that any mega developments will take place in the future or just happen, no earthquakes etc. I think that going without experts' opinions might still be very good because the trend is quite stable. [Academic] • ... using statistical methods may lead to a lack of some expert information. But still, I think statistical methods can tell you some of the logic behind those forecasts, whereas experts' opinions may be based on personal experience. • Statistical forecasting is more scientific in the sense that it is calculated by numbers and formulas rather than subjective opinions from experts. It is more straightforward if you know how to predict using statistical tools. [Academic]
Statistical and judgmental forecasting are complements.	• …you can interview expert panels and ask for the reasons behind their choices. I think this is something you cannot get from statistical methods. [Academic] • ... sometimes judgmental forecasts can improve forecasting performance because econometric or statistical models depend very much on the historical data, and so they can pick up the historical patterns using these models. But if there are a lot of uncertainties in the future, the model will not be able to anticipate these uncertainties. However, some experts may know about these uncertainties. [Academic]

Table 5.35 Summary of experts' opinions on integration (Continued)

Item	Examples of Comments
	• Because as I said, for statistical forecasts, it is just [a case of] learning from past experience to see the trend. There maybe some new features or new events, such as SARS and IVS. These cannot be projected by statistical forecasts. Also expert predictions are better than statistical forecasts. [Industry] • Some trends and patterns are observable in the statistical forecasts, for example, the GDP, the economic well-being of the country affecting the residents' desire to travel. Another factor is the exchange rates, and this is also a key factor [affecting demand]. It is measurable. Statistical models may do a better job than experts' views, but there are some preferences (e.g. consumers' preferences) you cannot measure from statistical models. [Industry] • Statistical methods may give you some logical results, and expert predictions may give you some insights into the industry, which cannot be obtained solely from statistical models. So a combination may capture the advantages of both two methods. [Academic] • Expert predictions may absorb some of the benefits of the (raw) statistical forecasts by running the models. [Industry] • So statistical forecasts are the foundation, and they should be concurrent with the experts' views. [Industry] • The pattern of seasonality is reflected in the quarterly statistical forecasts… I will see if it matches [my expectation] or if there are some patterns that are not observed in the forecasts, and then I may make adjustments. [Industry]
Select experts appropriately to reduce biases.	• … [if the] experts you have chosen do not have expertise or sufficient knowledge of the method or the subject of the topic, then probably the results will be biased. Largely on the basis of their intuition, and sometimes by luck, experts may produce accurate results. [Academic] • You want to get good ones, the ones who can really tell or predict. The number does not really matter. We want quality not quantity. [Academic] • Maybe you think that when experts have nothing of value to contribute, you should not combine, but it is a personal judgment. If we cannot find the right experts, we should not combine. [Academic] • But if you ask some experts, especially those experts who know forecasting techniques very well, they may be able to tell whether you have used the correct methodology, and if they think the methodology is incorrect, they may actually use their own expert skill and expertise to adjust the statistical forecast as they know that basically their domain knowledge is more useful than the forecasts themselves. [Academic]

Table 5.35 Summary of experts' opinions on integration (Continued)

Item	Examples of Comments
	• I will expect that the experts may have some statistical background or support for them to make predictions. [Industry] • Only if you select a pool of people who have great expertise and when I say expertise, I mean that these people know what they are doing and have been doing it for a while. Also they understand your forecast and why their forecasts are different from it. [Industry]
Examples provided by experts to illustrate how they made adjustments	• … when we were doing arrival forecasting for Singapore and according to the statistical methods, we were saying that Singapore arrivals would reach 1.6 million a year or so a year. The management challenged us, saying how about the population ratio: Does it make sense? As the population does not really grow in Singapore, how can arrivals be growing when the population does not grow that much? So, how did we justify? We used other different ways to verify our forecast; whether it made any sense or not is a business judgment. [Industry] • If we follow the statistical trend we have been through in the past 5-10 years, if that trend was to continue, we would be looking at 92 million visitors by the year 2020. Now consider the human side or the expert side. First of all, Hong Kong only has a certain amount of space, so is it realistic for us to say that we could actually welcome 92 million tourists in terms of buses, trains, and hotel rooms? …I do not think that we should just look at the statistical forecasts alone. I think we have to look at what people are thinking or feeling about them. … If you sat in your office and just looked at those statistical forecasts, you would think "Yes, that is what the model says, but physically, realistically, is it possible?" [Industry]
Recommendatio ns on statistical forecasting models	• … we should open our minds to see what additional [forecasting] methods can be used and more importantly, how accurately these methods could be applied in tourism demand forecasting. [Academic] • … from the hotel supply perspective, there may be constraints, such as how many visitors can visit Hong Kong at a certain period of time…. Statistical models need to take the capacity constraint into account. [Industry] • There maybe limitations to the statistical forecasts. You may need to find more variables and more sophisticated models to account for these measures. [Industry]

Despite some of the benefits of integration, the respondents also recognized that there are some drawbacks to statistical and judgmental forecasting methods. Two drawbacks of using human judgment in adjusting statistical forecasts were identified. The first disadvantage is the high reliance on the statistical forecasts. One expert expressed that he often made minor or no adjustments to statistical forecasts; instead, he tended to be too reliant on statistical forecasts in the adjustment process. The second drawback is that the involvement of human judgment introduces bias; for example, one expert stated that if the chosen experts did not have sufficient knowledge of the forecasting methods, or the subject of the topic, the forecasting results might be biased.

In addition, several respondents addressed the importance of selecting qualified experts to make adjustments. They generally agreed that qualified experts with appropriate expertise should be selected for the adjustment process. This is in accord with the findings from Kollwitz (2011), who disclosed that the forecasting results of a Delphi panel depend on the knowledge and cooperation of participants and thus it is essential to recruit participants who are likely to contribute valuable ideas.

However, there is no consensus on how to define “expertise” in order to select experts. One industry forecaster elaborated that a suitable expert is someone who “knows what he is doing and has been doing this for a while”; the other experts thought that an expert should not only understand the statistical forecasts provided but also understand why his/her forecast was different from the baseline forecasts (statistical forecasts). Moreover, this respondent also expressed his concern about the number of panellists included in the Delphi survey in this study − he believed that the greater the number of people included, the higher the possibility that the original forecasts would be diluted, particularly when these forecasts are of high accuracy. This expert further recommended that only those tourism forecasters who really have rich market knowledge and forecasting experience should be kept on the panel. This finding helps to clarify one possible fallacy of using the Delphi method, namely that “the more individuals are involved with a Delphi as users, the more effective it will be” (Linstone & Turoff, 2002, p. 569), and suggests that the quality of Delphi forecasting results depends on the quality of panellists rather than the number.

The ability of statistical models to correctly capture future trends was challenged by one industry expert who stated that historical trends would possibly not continue into the future all the time or at least not continue at the same pace. He further stated that “purely rely[ing] on the historical data or relatively short historical data to predict the future” often leads to underestimated results in tourism forecasting.

In terms of the reasons for integration, the prevailing view among the respondents was that judgmental and statistical forecasting methods are complements, suggesting that the integration of these two methods can absorb each method’s merits and overcome their shortcomings. Table 5.35 shows that under certain circumstances (e.g. series with stable trends), some of the respondents used statistical and judgmental forecasting methods as substitutes rather than complements. For example, one expert stated:

> *We do not predict that any unusual things will happen, that any mega developments will take place in the future or just happen, no earthquakes, etc. I think that going without experts’ opinions might still be very good because the trend is quite stable.*

(7) Assumptions of adjustments

Although forecasting is a key input for strategic management in the tourism industry, it is still in its infancy and a naive attitude towards forecasting can be observed (Kollwitz, 2011). This is reflected in the frequent acceptance of forecasts without challenging or even considering the underlying assumptions. A forecast may be accurate but based on wrong assumptions, while under other circumstances, a soundly reasoned forecast might not eventuate (Faulkner & Valerio, 1995; Kollwitz, 2011). Forecasters need accurate descriptions of the conditions underlying a forecasting problem in order to develop generalizations (Armstrong, 2001b).

This section assesses the underlying assumptions made by the experts during their forecasting adjustments in the Delphi surveys. The respondents possessed different ideas about what underlying assumptions lay behind their adjustments. The assumptions made by the experts were categorized into four factors (see Table 5.36): market knowledge and characteristics, tourist behaviour, statistical forecasts as the foundation, and feedback/comments from other experts. A more detailed summary is provided in Table 5.36.

The interviews showed that the majority of the experts considered market knowledge and characteristics, such as economic conditions, social situations, political issues, impact of foreseeable events in the future, tourism policies, transportation and accommodation, and other factors (e.g. promotional programmes), as a key assumption determining their adjustments. Among a list of market information, the experts regarded economic conditions as the most important factor to guide experts on how to make their adjustments. Some of the experts indicated that their adjustments were largely based on their personal beliefs about the markets, namely what they thought about the market and what was happening in the market.

Table 5.36 Key assumptions for experts' adjustments

Key assumptions	Factors
Market knowledge and characteristics (14)	• Economic factors (e.g. economic growth, business cycles, income level, exchange rates, inflation, employment, etc.) [14] • Social factors (e.g. levels, its development, distribution, and density of population growth) [1] • Political issues/factors [3] • Impacts of specific events in the near future (e.g. London Olympic Games, Japan earthquake, European crisis, economic crisis, the high speed railway, direct flights between China and Taiwan, etc.) [7] • Travel-related policies (e.g. visa policies, tax policies) [2] • Transportation and accommodation (e.g. availability of direct flights, flight capacity, variety and flexibility of flight options, hotel demand, hotel supply constraints, etc.) [4] • Other factors (e.g. promotional programmes and activities from destinations, destination images, etc.)
Tourist behaviour (5)	• Tourist characteristics, tourist preferences, consistency of tourist behaviour, convenience of travel, travel patterns, type of tourists, travel modes, purposes of visits, etc. [6]
Statistical forecasts as the foundation (3)	• I believe there is something missing from, or not fully captured by, the statistical forecasts. • … there are some factors in the future development or growth of tourism demand that [cannot] not been captured by statistical methods. • I look at statistical forecasting performance to see whether there are any obvious problems in the forecasting based on my own understanding of visitor arrivals trends in Hong Kong.
Feedback from other experts (2)	• I think I will adjust [my forecasts] after seeing other experts' opinions because maybe one's thinking is too extreme or there are some outlier figures or forecasts one should consider. It is good to know what others are thinking. • I will look at the data, the average adjustments, the reasons why these forecasts were made, and why these panel members have particularly high or low forecasts. There maybe some elements or reasons that I have not considered; after taking into consideration their concerns, I will adjust [my forecasts] for a second time.

Seven experts reported that they made their judgmental projections based on an assessment of the possible impacts of specific events (e.g. the earthquake in Japan in 2011, the launch of direct flights between Taiwan and the Mainland) over the forecasting period. One expert claimed that she relied highly on such events in making her adjustments. High reliance on using event information to make adjustments would probably bring the availability bias into the final adjusted forecasts. This type of bias has been commonly recorded in the judgmental adjustment studies where researchers have found that individuals are overly influenced by recent events or by events that are easily recalled (Önkal, Thomson, &

Pollock, 2007; Stekler, 2007).

In addition to the assumption of market knowledge, three experts stated that statistical forecasts also provided them with a solid foundation for their judgmental adjustments. They also believed that their judgmental inputs could help to overcome the drawbacks of statistical forecasting methods when they were not able to capture the impact of foreseeable events.

A few of the interviewed experts indicated that they had slightly changed their adjustment assumptions over the rounds of the Delphi survey. These adjustments were due to the availability of new market information or to feedback from other experts. For example, one expert made the following comment: “in the second round, I was more pessimistic because of the (European) financial crisis.” In addition to the assumptions made in the initial round, the experts who participated in the second round also considered feedback from other experts in revising their original adjustments. For example, one academic expert commented that:

> *I will look at the data, the average adjustments, the reasons why these forecasts [were made], why these panel members have particularly high or low forecasts. There are maybe some elements or reasons that I have not considered; after taking into consideration their concerns, I will adjust [my forecasts] for a second time.*

It is worth noting that all of the industry experts expressed a common concern about the growth of arrival forecasts based on supply constraints (e.g. shortages of hotel accommodation, passenger transportation capacity, and flight capacity). They doubted that statistical forecasts were too optimistic as forecasts from statistical models appeared to be “nonstop” and too robust.

The importance of including supply constraints into tourism demand models has already been addressed in the existing literature. For example, Bonham, Gangnes, and Zhou (2009) argued that in their study, demand parameters could not be estimated reliably without regard to supply constraints and potential price responses because the two markets (USA and Japan) had a dominant share (85%) of the total market. According to the visitor arrival statistics released by HKTB (2011), the six source markets (Mainland China, Japan, Taiwan, Australia, the UK, and the USA) examined in this study represented a dominant 81 per cent of the total inbound market in Hong Kong in 2011. Forecasts of total inbound visitor volume based on this remarkable market share would probably be misleading if the supply constraints were not well taken into account.

(8) Use of information to make adjustments

The research on judgmental forecasting shows that the type of heuristic that people use to make their forecasts is largely driven by the type of information upon which forecasts are based (Harvey, 2007). Leitner and Leopold-Wildburger (2011) found that several sources of information, such as past sales or forecasts of other departments, are used by managers to forecast sales. This section explores the types of information that the respondents utilized when making their judgmental adjustments during the Delphi surveys and discusses the heuristic strategies they used and the major types of biases that arose from the applied heuristic.

It was found that the experts used multiple pieces of information instead of relying on one single piece of information. Their adjustment decisions were made based upon a combination of different assumptions and factors concerning individual source markets. One industry expert stated that his organization used multiple sources of information to obtain its final forecasts. This organization compared forecasts from external organizations such as PATA, UNWTO, and Euromonitor to aid its forecasting decisions. The following quote provides an example with regard to the information considered by one expert in her organization's forecasting decisions:

> *We look at the historical information. We look at the rail capacity. We look at the population, the disposable income. How many people could go there? How many people could travel there? How many people are substituting flying for rail? … We have a very robust travel trade network. So we can ask our trade partners how much they are selling. We also have data on hotel rooms, and we can benchmark [our forecasts] against this data. There is no model telling you exactly to what extent human judgment should be involved in making final forecasts, but there is a sense that can tell you that how much it should be used.*

As revealed in the previous paragraph, the experts made their forecasts on the basis of different types of information. Harvey (2007) described three types of information and their corresponding heuristics: the availability heuristic is employed when forecasts are made from "information held in memory", the representativeness heuristic is used when the value of one variable can be forecast from explicit information about the value of another variable, and the anchoring-and-adjustment heuristic is applied when the value of a variable can be predicted from explicitly available information about previous values of that same variable.

The three types of information described in Harvey's (2007) study (past information about the forecast variable, information held in memory, and information about the variable being forecast and its influencing factors) were also used by our Delphi experts when making their adjustments (see Table 5.37). In other words, three heuristics – anchoring-and-adjustment, availability, and representativeness – were employed by the experts during their forecasting adjustment process. According to our interview results, the experts did not employ just one heuristic to make adjustments but rather a combination of heuristics. Research into the heuristics and biases has suggested that heuristics do not provide an optimal way to make judgments and decisions; rather, they can sometimes lead to systematic errors or biases (Harvey, 2007). In the following discussions, we make connections between a specific heuristic and the type(s) of information used and discuss the possible biases that were revealed either in the interview data or by the quantitative analysis presented in Section 5.5.3 of Chapter 5.

Table 5.37 Use of heuristics and types of information to make adjustments

Type	Examples of comments
Anchoring and adjustment [14]	Past information about the forecast variable: *Historical data of the arrival series* • I guess I did refer to historical data, as I cannot do it without this data. • That is the whole thing about forecasting − we look at the past and predict the future. • I did check the historical data. • I think there are two major concerns in my predictions: one is the economic factor, the other is the historical trend. • I will consider the historical figures, mostly the patterns. • ... showing the historical trends which you used for statistical forecasting can also give information or an indication to experts to consider whether the historical trend is too low which affects your statistical forecasts, and therefore requires some expert adjustment.
Availability [14]	Information held in memory: knowledge and personal beliefs about the market, knowledge and characteristics, tourist behaviour, impacts of recent events, etc.
Representativeness [14]	Information about the variables influencing arrivals: • ... they [tourism forecasters] check the GDP, other economic indicators, or other official sources to make their forecast[s]. • We look at the economic data. This type of data is obviously vitally important because if a particular country which provides visitors is not doing well, then people from that country are not going to spend money coming to Hong Kong. If a country is experiencing growth, then there is a good chance that we can look upon it much more positively. If a country is not experiencing growth, it is either going [in]to recession or getting ready to go [in]to recession, and so we have less optimism about it. • We rely on transportation as a major forecasting factor in bringing people to Hong Kong. • We look at the trends: the business trend and the travel trend.

The role of anchoring and adjustment

The extrapolation of historical information about a variable is probably the most common way of conducting forecasting in applied contexts (Harvey, 2007). Harvey (2007) observed that people often use the last data point of a data series as a "mental anchor" and then make an adjustment based on that anchor to allow for the consideration of the major features of the series. Wright and Ayton (1987) showed that the provision of past information is likely to have a positive impact on forecasting performance "where judgmental forecasts are performed on a repetitive and sequential basis" (p. 113).

In the Delphi survey in this study, the experts were provided with a full vision of the historical data (in graphical format, see Appendix I of Appendix A) and recent 4-year data (in graphical and tabular format, see Appendix A) 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. For example, the following quote shows that 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.*

In terms of the length of the historical data presented to the experts, four of the nine academic experts suggested "*the longer, the better*." Some experts further explained that it was quite difficult to foresee future trends by only reviewing 4 years of data (as provided by the HKTDFS); they stated that it was not sufficiently long enough to observe historical patterns. The provision of full historical data can enable forecasters to gain additional information about the raw data and forecasts; for example, a longer span of historical data can better help forecasters' adjustment decisions by allowing them to judge whether a forecasting trend is too low or too high. The examination of historical data can help forecasters to find out what factors are ignored or missed by statistical forecasts.

The interview results showed that the experts held different views on what constitutes an appropriate length of historical data. Some experts suggested a historic series of 10 to 15 years would be most helpful, while others believed that some rules of thumb should be used as guidance in order to select the length of historical data in comparison to the length of the forecasting period. For example, one expert stated that "every 5-7 point[s] should be used for one prediction point". However, another academic expert argued that the longer the historical data provided, the greater the possibility that the forecaster could be misled because it is possible that old historical data may not be representative of current trends or the current growth of tourism demand.

In contrast to the academic interviewees' views on the length of historical data that should be provided, most (4 out of 5) of the industry experts relied more on recent data than past data. They believed that recent demand observations contained more information about the future than past observations. Three of the five industry respondents thought that recent 4-year data was sufficient for their adjustments. One of the respondents reported that she actually relied more on recent quarters (2-3 years), and she believed that it was not necessary to go back as far as 4 years ago. Only one expert stated that 4-year data is too short because it is not able to capture trends and patterns, especially for seasonality and some ad hoc events in certain years.

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 checked both the historical data and the statistical forecasts in order to make reasonable adjustments. There was a strong majority who declared themselves anchored around the statistical forecasts provided, and they only adjusted the forecasts if they believed that it was absolutely necessary.

The results presented in Chapter 5 show that the judgmental forecasts in this study were, on average, unbiased for the six source markets. However, evidence of bias was found in predicting arrivals from Australia, China, Taiwan, the UK, and the USA. A mix of different trend patterns possibly cancels out the effects of bias from individual markets. It is necessary to investigate the reasons for the causes of biases in order to reduce bias in future judgmental adjustment forecasting.

As discussed previously in this section, the experts were found to be anchored on the historical trends or the forecasts provided to them. Harvey (2007) suggested that adjustments based on anchoring the last data point are typically insufficient in that the “forecasts are below the optimum for upward trends but above the optimum for downward ones” (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.

Many studies have shown that when people make initial forecasts and then take advice from a more expert source such as a decision aid, an expert system or a superior judge, they do not make full use of that advice (Harvey, 2007). In our interviews, it was also found that the experts did not take sufficient account of the comments/feedback from other experts. One expert reported that she believed that her judgments were right and so she kept her forecasts unchanged in the second round. Two other experts indicated that they checked other experts' comments but would selectively choose useful information; however, they did not elaborate on how they distinguished between useful and irrelevant information. Some experts reported that they found other experts' advice and comments very useful because they may have ignored some important information that could affect the variables to be forecast. Two tourism officers explained that in their case, underestimation was partly due to organizational pressure.

Krueger (2003 cited in Harvey, 2007) disagreed that the insufficient use of advice is due to underadjustment from an earlier anchor; rather, he distinguished the effects of anchoring and conservatism: conservatism is shown by the way people are reluctant to change their opinions based on others' advice because they regard their own opinions as the correct ones, whereas the anchor is provided by the data rather than by the forecasters (Harvey, 2007).

The role of availability

The use of the availability heuristic suggests that people should consider information that is more easily retrieved from memory to be associated with more likely events (Harvey, 2007). Gigerenzer (1996, cited in Harvey, 2007) argued that availability is not specified sufficiently to make quantitative forecasts. The discussions in the section below explore how the experts in this study used the availability heuristic in forecasting from information held in memory.

The results from testing the hypothesis on the effects of contextual knowledge on accuracy in Chapter 5 showed that the industry experts produced statistically more accurate forecasts than the academic experts. 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. Japan earthquake, 2012 London Olympic Games) by making adjustments to the initial forecasts. By contrast, the industry experts relied greatly on their knowledge and forecasting experience with regard to the market concerned. Their insights into the market place are a combination of their own understanding of the market and views from their network partners. This approach helps industry experts to obtain a holistic view of the tourism industry they are working in which will then help with their forecasting decisions.

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.

As shown in Table 5.7, another expert commented that the group forecasts in the second round were too low because “although the earthquake will affect the country’s income level in the short run, Japan will have new opportunities to rebuild the economy, and economic growth will increase as a result.” If all of the information about the earthquake had been used, it would have suggested that the negative impact of the earthquake would be temporary and that the market would return to its long-term equilibrium.

Furthermore, one academic respondent believed that the more information given to experts, the more confident experts could be in analysing and adjusting for the future. As summarized by Makridakis, Wheelwright, and Hyndman (1998), the above belief is categorized as one type of conventional wisdom which can threaten decision-making effectiveness. Empirical findings suggest that the amount of information does not improve the accuracy of decisions; instead, more information merely seems to increase people's confidence that their decisions will be correct without necessarily improving the accuracy of their decisions (Makridakis, Wheelwright, & Hyndman, 1998).

The role of representativeness

According to the representativeness heuristic, the value of one variable can be predicted from the value of another variable. Our interviews showed that when predicting arrivals from a specific market, the experts utilized several sources of information to reach their adjustment decision. Two examples are provided below:

> *... Take Singapore for example. You will say, Singapore currently has [a population] about 500,000 a year, its economy is good, its GDP growth is good, all of its people are employed, there is no negative employment, all Singaporeans have spending power and have a desire to travel, etc. Then we will say, all of that data tells us that Singapore could probably grow by another 5 per cent. So, we look at this information as a way of forecasting. We follow the same logic for each of our markets, and that is how we automatically come up with our forecasts. [Expert 1]*

> *We have a very robust travel trade network. So we can ask our trade partners how much they are selling: Are you selling more this year? We also have data on hotel rooms, and we can benchmark [our forecasts] against this data. There is no model telling you exactly to what extent human judgment should be involved in making final forecasts, but there is a sense that can tell you that how much it should be used. [Expert 2]*

The two examples above illustrate how two respondents made their forecasts by considering a number of factors in their forecasting. The first expert considered macroeconomic and social factors such as economic growth, employment, the population of the source market, the spending power of consumers in the source market, and the travel desire. The second expert focused on supply chain information sharing. Clearly, the experts did not use algorithms to combine all relevant information into their forecasts but rather used their mental strategies to produce their final forecasts. However, forecasts based on this mental prediction rule may not be maximally representative of the input information.

The experts also reported that they found the IMF projections on GDP and exchange rates, provided during the Delphi survey useful in helping them with their adjustment decisions. Some experts stated that they also searched for any relevant information on the Internet to assist with their adjustments. To recap, the experts represented all aspects of relevant information in their forecasts by using the representativeness heuristic.

(9) Data presentation format

Armstrong (2001e) reported that one aspect of the structuring process that can affect forecast accuracy is the presentation format of the data. Past research has suggested that in some cases, graphical presentations improve, and in other cases harm, forecast accuracy. For example, Harvey and Bolger (1996) concluded that graphs led to more accurate judgmental forecasts for series containing trends and tables led to more accurate forecasts in other cases. They also stated that graphs of trended data could help forecasters avoid underestimating the steepness of trends, but on the other hand, graphs seemed to promote inconsistency and overforecasting bias for series without clear trends.

To explore the experts' data-presentation format preferences, the experts were required to answer a question on what display format they used when making their adjustments in the HKTDFS. It was found that the experts used graphs and tables differently, and held different views about the influences on their forecasting adjustment process.

Some of the experts stated that they only referred to graphs during their adjustments; for example, one expert asserted that she preferred to look at the graphs because they provided a clearer way to observe fluctuations, trends, and the gap between predictions and adjustments. Similar comments were reported among the respondents who claimed that they only used graphs during the adjustment process.

In contrast, other experts believed that tables were more important and informative than graphs; for example, one industry respondent explained that he relied more on tables due to a lack of understanding of the graphs as provided. It is interesting to note that the respondents who preferred tables were all industry practitioners. This is probably because tourism practitioners are more concerned with changes in the number of arrivals (either in absolute or relative terms) in the short term due to the peculiar nature of the tourism industry (i.e. full of uncertainty and dynamics, volatile, sensitive to external shocks, etc.).

Although the experts had different preferences regarding the use of graphs or tables to adjust statistical forecasts, the majority of the experts agreed that it was more straightforward and easier to check and use graphs. For this reason, the industry experts tended to rely on graphs more than tables. Both the academic and industry experts agreed that the provision of both graphical and tabular forms was helpful and useful for their adjustments. One expert elaborated that tables can provide more detailed information in terms of numbers whereas graphs are useful in terms of visual effects.

(10) Difficulty of forecasting tasks

During the interviews, the experts were required to indicate which market(s) they thought were more difficult to predict compared to others. The majority of the respondents agreed that the Chinese market was the most difficult to forecast– nine out of the 14 experts rated this market as the most challenging one to forecast. Table 5.38 shows that there were two important factors that the experts considered in relation to this market: it is a policy-driven market and it is an emerging market with full dynamics and complexity of tourists' behaviour. They also suggested that in order to improve forecast accuracy for the Mainland market, it should be investigated at the subregional level rather than the national level; for example, forecasts could be disaggregated into province level or even city level. One industry expert proposed that forecasts could be made by distinguishing IVS and non-IVS visitors.

With regard to the other five source markets (i.e. Japan, Taiwan, Australian, the UK, and the USA), the experts held different views on the difficulty of forecasting visitor arrivals from these markets. Some of the respondents stated that it was easier for them to make adjustments for markets they were familiar with or that were closer to Hong Kong as they had more market knowledge and felt more confident to make adjustments, as illustrated in the following comments made by two experts:

> *I am more familiar with [the Australian market] because I read more and hear more about it, I have lived there before, I have friends there, and I have visited these places more. [Expert 1]*
>
> *I know more about this market, such as its culture, development stages, and the like. [Expert 2]*

Table 5.38 Experts' comments on the difficulty of forecasting tasks

Market	Examples of Comments
China [1st: 9; 3rd:1]	• **Volatile**. … I think [that with regard to] the China market, the information, the dependent variables, and independent variables **fluctuate** so much. … The China market is so **diverse** in terms of the tourism demand from its first-, second- and third-tier cities. I think all of these trends are very different. The demand from first-tier cities might be declining, but demand from second-tier cities might be rising by 20%, whereas demand from third-tier cities may only rise by 5%. The rise in demand from people at the lower-income level might be a little slower. People with a very high income who have travelled to Hong Kong ten times before have probably shifted their interest away from Hong Kong. [Academic] • … this is an **emerging** market. [Academic] • There is a lot of **uncertainty**. This is largely **policy-driven**. … many people come to Hong Kong to **give birth** to their children. … They come here not for the purpose of travel but for many other purposes, such as investment and buying apartments. I use the word, **unusual travel behaviour**. … That particular source market exhibits very different consumer behaviour compared to other markets. [Academic] • Because the influences are not just the economic situation in China but also **policy issues**. [Academic, Industry] • **Political** things can happen in a minute. [Industry] • We have noticed that the trend is going up somehow, but we just do not know how robust this could be. But other [markets] are pretty stable. [Industry] • China is the most **challenging** market. First, [it] is **changing very fast**; second, [it] is **growing very robustly** and all growth seems to be **nonstop**; third, [it] is very **dynamic**. [Industry]
Japan [1st: 1; 2nd: 3]	• The impact of the **earthquake** that occurred in **2011**. [Academic, Industry] • Because according to the [historical] studies in the past, with regard to demand for Hong Kong tourism from Japanese tourists, sometimes it does **not comply with the economic theory** and sometimes the **income variable has been insignificant** in past studies. [Academic]
Short-haul markets	• I think the ones [markets] **closer to** us **within the region** are probably **easier** to predict. [Industry]
Australia [2nd: 1]	• **Less familiar** with the market. [Academic]
UK [1st: 2; 2nd: 1]	• **Less familiar** with the market. [Academic] • …because the historical trend of visitor arrivals, i.e. some different patterns in the history, makes it **difficult** for me to make predictions. [Academic] • Of them all, the most difficult one is probably the UK. … Well, this is because of the volatility. I do not know exactly which way things will go. Maybe nothing will happen. [Industry]
USA [1st: ; 2nd: 2]	• **Less familiar** with the market. [Academic]
Long-haul markets	• … maybe I do not have much knowledge about these long-haul markets. … Also, factors affecting long-haul markets may be due to the price of air tickets and the time issue. For long-haul travel, it takes much longer time to plan and make a trip, but tourists who go to short-haul destinations make travel decisions for a very short time. There may be more factors that affect long-term decisions. [Industry] • I think it is easier to predict mature markets, such as European and US markets. [Industry] • These markets so mature [Industry]

From the perspective of market conditions, one academic expert thought that markets where there is not much change in the environment are much easier to predict. She further elaborated that if there were no dramatic changes in an environment, she tended to believe that the arrival forecast would stay relatively stable over the forecasting period.

The experts had conflicting opinions regarding the difficulty of predicting long-haul markets. One industry respondent indicated that mature markets, namely long-haul markets, were easier to predict, while other academic experts held the contrasting view – long-haul markets were difficult to predict because they did not have sufficient market knowledge to make forecasts.

(11) Useful features of the HKTDFS in assisting the experts' adjustments

The results from the hypothesis tests show that judgmentally adjusted forecasts on the basis of statistical forecasts are more accurate than statistical forecasts alone. This section examines possible reasons for the accuracy improvement with a particular focus on the features of the HKTDFS.

All of the respondents participated in the main Delphi survey conducted at an earlier stage. As shown in Chapter 3, they were given access to the HKTDFS and presented with system-generated statistical forecasts, and then they were required to make their adjustments to the statistical forecasts and provide their reasons for making adjustments. A number of factors were identified while asking the experts to recall the features of the HKTDFS that were most useful in aiding them to make their adjustments. There was general agreement among the respondents that the HKTDFS is a user-friendly system (see Table 5.39).

The forecasting tools in the HKTDFS greatly supported the experts' forecasting adjustment process. One expert who had experience with both spreadsheet forecasting and computer programs described her efforts with the HKTDFS as being more "substantive" and "rewarding" than the efforts associated with the manual data input that is necessary when forecasting in Excel. However, it became clear from the interview data that without the proper tools and technology, effectively organizing, storing, retrieving, and analysing abundant sources of data can become a daunting task leading to the underutilization of data. The analysis of the interview data suggested that the system may not be sufficiently effective in enhancing a forecaster's performance if he/she lacks knowledge and experience of using the information system to assist with the forecasting process. For example, one industry expert relied greatly on his "gut feelings" and usually used "pencil and paper" to make his forecasts. An examination of this expert's individual forecast accuracy showed that his forecast accuracy was the poorest among the industry experts involved in the Delphi surveys, which confirmed the above presumption.

Table 5.39 Key factors contributing to experts' adjustments

The system is user-friendly. [11] - Easy clicks. - It is easy to use. - I think it is pretty user-friendly. - In general, it is quite easy. - I think the ease of using the forecasting system is very crucial. - The system is also well-designed, and it is easy to understand what is going on. - It is quite good. It is easy to use. It provides the essential information to help me to do the forecasts. - The interface of the system is nice-done. - The system is user-friendly. - I think first thing, it is very user-friendly. - I think it is very good to use the system.	**Decisional guidance:** - Graphs or tables of time series. [9] - Provision of reason feedback. [7] - Provision of historical data. [6] - Provision of statistical forecasts as baseline forecasts. [4] - GDP growth rates and exchange rates forecasts from IMF. [4] - Event information provided in the instructions. [3] - Provision of the 1st round's summarized forecasts in the 2nd round. [2] - Ask for verification of adjustments.[2] - Instructions to make adjustments. [1]
Other factors: - Two adjustment options (by annual or quarterly adjustments). [3] - Use of Delphi technique. [1]	

The interview data identified multiple types of information guidance that were used to motivate the forecasters to add value to the forecasting adjustment process. The majority of the experts reported that the information guidance summarized from the interview data was of great use to them during their adjustment process (see Table 5.39). The graphs and tables of the series, the provision of statistical and group forecasts as baseline forecasts, and the provision of event information over the forecasting period were particularly useful. Several respondents (9 of 14) highlighted their reliance on graphs and/or tables of time series to assist their adjustments.

Most of the experts stated they checked the opinions of other experts in the second round to see if they had overlooked some important determinants or if they had made an adjustment for the same reason but of different magnitude. This is consistent with the findings from Goodwin and Wright (2010), who commented that the exchange of reasons between panellists could alert them to "inappropriate framings, biases in the recall of similar cases, utilization of inappropriate reference classes, cognitive bias, and inappropriate views of causality underpinning the unfolding of event chains" (p. 362). Some of the experts suggested that the system could be improved if it offered built-in guidance to provide information regarding the purpose(s) of a Delphi survey, a list of influencing factors affecting forecasting, and a brief interpretation of the graphs.

(12) Recommendations on improving the HKTDFS

Our discussions have mostly been concerned with how the HKTDFS was used. During the interviews, suggestions and recommendations were obtained from the experts on how to improve the system to make it more effective. Fildes, Goodwin, and Lawrence (2006) summarized that there are major gaps in the knowledge of how an FSS should be designed to support its users' decisions about when and how to intervene judgmentally. They also suggested that guidance is likely to be more productive than restrictiveness. A forecasting support system with higher restrictiveness is likely to be much easier and simpler to use, which is more acceptable to many poorly trained users, but on the other hand may forgo the potential benefits of human involvement in forecasting and therefore reduce forecast accuracy.

However, Fildes, Goodwin, and Lawrence (2006) also identified the propensity of support system users to over rely on judgment and the failure of certain forms of guidance to curb this tendency. Past studies have demonstrated that it is possible to gain benefits if users of an FSS develop a sense of ownership of the forecasting outputs and if an FSS enables the role of explanations in selecting a suitable forecasting method. The fundamental determinant of result demonstrability (defined as the "tangibility of the results of using the innovation") (p. 358) in an FSS is likely to be the perceived accuracy of the forecasts (Fildes, Goodwin, & Lawrence, 2006).

A summary of the various recommendations made by the respondents is presented in Table 5.40. The recommendations made by the respondents focused on two aspects, one regarding how to relax the restrictiveness of the system, the other concerning the provision of decisional guidance to help judgmental inputs. The findings with respect to decisional guidance were summarized according to three design strategies (informative, suggestive, and predefined guidance) based on Fazlollahi, Parikh, and Verma's (1995) study.

Table 5.40 Recommendations on improving the effectiveness of the HKTDFS

<table>
<tr><td colspan="2">Restrictiveness</td></tr>
<tr><td>(1) Data & variables</td><td>(2) Decomposition</td></tr>
<tr><td>- Provide full historical data.
- Provide annual and quarterly growth rates of raw data and forecasts to supplement the absolute numbers in the current system.
- Allow forecasting series to be adjusted in a more interactive way to incorporate experts' judgments of forthcoming events in the forecast period.
- Add additional explanatory variables in the forecasting models, e.g. fuel price, marketing expenditure.</td><td>- To increase the accuracy of predicting the Mainland market, it is better to have a breakdown of visitor arrivals by subregion (e.g. South/Central/East/North China), by cities, or visiting schemes (e.g. IVS and non-IVS).
- Enable adjustments at both subregional and aggregated levels (particularly for the Mainland market).
- Enable adjustments by travel patterns or tourist characteristics (e.g. same-day vs. overnight visitors, first-timers vs. repeat visitors, business vs. leisure visitors).</td></tr>
<tr><td>(3) Adjustment options</td><td>(4) Easy-to-use design facility</td></tr>
<tr><td>- Provide options to change intercepts, and the trend slope.
- Options for adjusting explanatory variables.
- Add options for more source markets.</td><td>- Allow judgmental adjustments to be made by drag and drop movements onto the graph to improve the system's user friendliness.
- Provide customized reports that can be exported and saved in Excel or PDF file format.
- Provide options to export and save raw data and forecasts.</td></tr>
<tr><td colspan="2">Decisional guidance</td></tr>
<tr><td>(1) Informative guidance</td><td>(2) Suggestive guidance</td></tr>
<tr><td rowspan="3">- Provide a brief and short summary for statistical and judgmental forecasts.
- Provide multiple groups of baseline forecasts produced by different statistical forecasting methods. A brief explanation of each method and an indication of the circumstances where its application is most appropriate should also be provided.
- Display the underlying assumptions of statistical forecasts for users.
- Provide feedback to users on the accuracy of their forecasts, both the statistical forecasts and the judgmentally adjusted forecasts.</td><td>- Rank the accuracy of baseline and benchmark forecasting models and suggest that the user choose a proper model with least forecast error.</td></tr>
<tr><td>(3) Predefined guidance</td></tr>
<tr><td>- Ask for a verification of adjustments, especially when very large adjustments are made or when there is a huge difference between a participant's adjustment and the statistical forecasts. A limit should be set in the system (e.g. 30%).
- Use provoking messages in order to disincline users from making continuous and maybe unnecessary adjustments.</td></tr>
</table>

Suggestive guidance is defined as the strategy used to make judgmental recommendations (what to do next, which input values to use) to the system users, thus influencing their decisions regarding choice of alternatives (Fazlollahi, Parikh,

& Verma, 1995). Informative guidance provides pertinent information to the user without suggesting how to act. This design strategy offers different alternatives without recommending which one is the best. Different from the above two guidance strategies, the predefined guidance strategy, which is prebuilt into the system and is rule-based, can be activated when required. The contents of the guidance, information, messages, or recommendations are predefined by the system designer. Such guidance is applied in a context-sensitive manner involving both the tasks and the history of user interactions.

The comments made by the experts regarding restrictiveness were arranged into four categories: data and variables, decomposition, adjustment options, and easy-to-use design facilities. According to the experts' comments, one problem with the HKTDFS is its limited access to the full historical data; only data of the past 4 years were accessible to the panellists. Most of the experts suggested that a longer span of historical data (e.g. 5 to 10 years) would be preferable.

The study by Fildes, Goodwin, and Lawrence (2006) indicated that decomposition is a process that can be incorporated into a system when a user wishes to make judgmental adjustments in which the underlying assumption is that a set of decomposed judgments are more accurate than a single holistic judgment. Comments on decomposition were mainly made with regard to the Mainland market. All of the experts interviewed believed that this is a special and vital market to Hong Kong and deserves special attention and additional efforts to improve forecast accuracy. Specifically, the forecast accuracy in predicting arrivals from the Mainland is likely to be improved if the judgmental tasks are broken down into a series of easier tasks, which would enable forecasters to take more information into account and thus produce more accurate forecasts. The experts suggested that judgmental adjustments could be made by breaking total arrivals from the Mainland into different regions (e.g. Guangdong and non-Guangdong areas), or different visiting schemes (e.g. IVS or group tours), or different types of visitors (same-day and overnight visitors, business and leisure visitors) and then aggregating these individual forecasts.

During the Delphi survey, the experts were only allowed to make adjustments to arrivals forecasts and only had one option in terms of baseline forecasts (i.e. econometric forecasts made by the ARDL-ECM approach). The respondents suggested that the HKTDFS could offer more adjustment options to increase its flexibility, such as the provision of options for more source markets, changing the model parameters (e.g. intercept, slope), and for adjusting explanatory variables.

All of the features and options within the system should be easy to use and understand. Although the experts acknowledged that the system is user-friendly, there is still some room for further improvement. One expert suggested that a judgmental adjustment could be made by dragging the data points to any position determined by forecasters. Another expert thought that it might be better if the system could provide an export option to download raw data and forecasts.

The features discussed in the previous paragraphs are concerned with how to reduce the restrictiveness of the system. The experts also made comments about establishing an effective guidance system in the HKTDFS. With appropriate guidance, flexible systems with many features can be made easier to use. The experts' suggestions on the provision of guidance were categorized as suggestive, informative, and predefined guidance. Two industry experts believed that it would be better if the system could provide them with feedback on the accuracy of both the statistical forecasts and judgmentally adjusted forecasts they have made in the past using the HKTDFS. Actually, Rowe and Wright (1996) concluded that the accuracy of judgmental forecasts can be greatly improved if timely feedback is supplied to panel experts. The other two academic experts suggested that the system could provide a group of statistical forecasts as baseline forecasts and briefly explain the use of each method for the experts' consideration.

Fildes, Goodwin, and Lawrence (2006) concluded that result demonstrability could be achieved through guidance in the form of clear explanations of forecasts and accuracy measures. Goodwin (2000b) found that users make fewer unnecessary judgmental adjustments when they have to provide reasons for their judgmental interventions. It is thus useful to provide information statistics and information about users' forecast accuracy in their past forecasting in the HKTDFS.

One industry expert also pointed out the importance of communication between academic researchers and industry practitioners because these two groups have different mindsets when forecasting. Another industry expert suggested that the system should avoid the display of redundant technical information as much as possible. However, an academic respondent expected the system to provide some diagnostic statistics to check the validity of the statistical forecasting models. Three industry experts suggested that tourism forecasters should hold regular meetings (e.g. focus groups or workshops) to exchange their ideas and views on tourism forecasting in Hong Kong. To design an effective forecasting support system that can truly reflect the needs of industry practitioners, communication is critical.

In addition to the experts' suggestions, the existing literature has also proposed a number of desirable strategies to make a forecasting support system easy to use. Fildes and Beard (1992) recommended that several facilities should be developed to obtain appropriate quantitative forecasts. For example, they suggested that a system should have an experimental module to allow for comparison among different forecasting methods. They also suggested that a system should allow the database to be split easily to enable postsample evaluation of the forecasting models and provide the ability to identify series where judgmental adjustment is likely to be appropriate.

(13) Contributions of this study: Experts' views

At the end of each interview, the experts were asked to list the contributions of this project from their perspective. The respondents agreed that this project would make valuable contributions to the tourism industry. First, this study provides reliable and accurate arrivals forecasts to depict the future tourism-demand trend in Hong Kong from different source markets. The importance of forecasting in the tourism industry was addressed by one respondent who stated that "forecasting is the scientific foundation for strategic decision making in many organizations, such as government policy making, tourism authority marketing initiatives, private sector investments, and operation management." It is therefore of great importance to provide reliable and accurate forecasts as a solid foundation for strategic decision-making.

Second, the uniqueness of the study lies in two features that were described by two experts: "this is the first forecasting exercise in the tourism forecasting field that has integrated statistical and judgmental forecasting methods together"; "it is the only forecasting system for the Hong Kong tourism industry that I am aware of." Other experts indicated that it might be a good idea to extend the application of the same integrative method to forecasting other tourism demand measures.

It was also interesting to find that the industry and academic experts had different focuses when addressing the contributions of the current study. The industry respondents emphasized the importance of interactive communication and collaboration between tourism researchers and industry practitioners in Hong Kong. Reliable and accurate forecasts were the top priority for the industry forecasters. However, one should not only rely on statistical forecasts alone, as one industry forecaster stated. Relying solely on statistical forecasts in predicting tourism demand does not ensure high accuracy and instead suggests that experts' intervention should be incorporated into final forecasts. The industry experts agreed that involving a pool of different experts to share their knowledge and experience is a good strategy. One expert shared his thoughts:

> *It is kind of sharing of thoughts. It can make our judgment better. It is also kind of how to make an accurate forecast because we are not looking at the issue from one side but more from different angles. Sometimes we do not know what is right: How can we know the highest speed of the railway? Maybe the government will know more than we do. Maybe involving them will be better for forecasting than simply including us*.

The academic experts appeared to evaluate and address the values of the current project in a more theoretical manner. For example, one respondent believed that the most innovative aspect of this study is the integration of forecasts obtained from advanced econometric methods with expert judgments within the tourism forecasting context. Another expert made the comment, from a methodological perspective, that "[the] methodology used in this study to forecast tourism demand is creative"; he believed that the results from this study would make significant contributions to the existing tourism demand forecasting literature. This expert further explained that social, cultural, and political issues have been overlooked in the tourism demand literature as it is difficult to measure them in numbers. As indicated by this expert, this study will provide insights for future tourism demand forecast research because it has employed human judgments from experts and integrated these noneconomic factors into advanced econometric models to forecast tourism demand. Another academic respondent believed that this study will provide some useful insights and valuable suggestions/recommendations for tourism researchers and practitioners on establishing an effective, efficient, and user-friendly tourism demand forecasting support system.

5.7 Chapter Summary

This study adopted a mixed mode of quantitative and qualitative analysis to evaluate the forecasting performance of statistical and judgmentally adjusted forecasts. To ensure reasonably good statistical forecasts, an econometric analysis, which included unit root tests, ARDL bound tests, and diagnostic tests, was conducted. Prior to the evaluation results, the arrival forecasts for the six source markets (Mainland China, Japan, Taiwan, Australia, the UK, and the USA) obtained from the Delphi surveys were presented. The forecasting performance of the statistical and judgmentally adjusted forecasts was evaluated from three dimensions: accuracy, bias, and efficiency. The research hypotheses were tested by examining the values of the error measures, conducting correlation and regression analyses, and employing statistical tests (both parametric and nonparametric depending on the test results of normality and the homogeneity of variances). Comparisons were made to examine the accuracy difference among different Delphi rounds, source markets, expert groups, expertise levels, levels of data variability, forecasting horizons, and sizes and directions of adjustments.

Compared to quantitative forecasting models, judgmentally adjustments to statistical forecasts cannot only draw upon inside information and experts' views about forthcoming changes and their implications based on their past experience but can also incorporate the up-to-date information that could not be instantly updated into the quantitative models. On average, statistical forecasts adjusted by the experts improved forecast accuracy for all of the six markets. The results showed that consensus group forecasts in the final round of the Delphi survey provided significantly more accurate forecasts than those of the initial statistical forecasts and the simple average of individual experts' forecasts in Round 1. Although satisfactory accuracy was achieved, the forecasts were found to be inefficient and biased for some of the individual markets.

After a systematic evaluation of the statistical forecasts and judgmentally adjusted forecasts, in-depth interviews were conducted to provide qualitative input to interpret the quantitative findings from the hypothesis tests, examine the underlying rationale embodied in the experts' forecasting adjustment process, and collect experts' opinions regarding the use of the forecasting system to aid their judgmental adjustments. The findings from the interviews confirmed that compared to the academic experts, the industry experts preferred to use simpler and easier forecasting methods. The experts reached the consensus that given the relative strengths and weaknesses of judgmental and statistical forecasting methods, it is necessary to integrate these two types of forecasts in order to make better tourism demand forecasts. According to the experts interviewed, a variety of reasons account for the great improvement in accuracy in this study, such as the provision of multiple information cues (e.g. time-series information and non-time series cues), the use of a Web-based forecasting support system, and the use of the Delphi technique to structure and aggregate experts' judgments. Useful recommendations and suggestions were made by the experts to further improve the HKTDFS and to point to future research directions.

Chapter 6 : Conclusions and Future Research Directions

6.1 Introduction

Sound accurate tourism demand forecasts can help tourism marketers, managers, planners, and others in public agencies reduce the risks of their decisions and the costs of attracting and serving the travelling public. One big challenge in making accurate forecasts is to utilize the best aspects of statistical forecasts while exploiting the value of human knowledge, experience and inside information about the market environment. Statistical forecasting methods allow for the extrapolation of established patterns and/or existing relationships in order to predict their continuation, assuming that such patterns/relationships will remain unchanged over the forecasting period. However, tourism demand forecasting, like the activity itself, is a diverse, dynamic, and changeable process that rewards quick and observant actions. Whenever changes are detected, or if changes are about to occur, human judgment is the only viable alternative for forecasting the possible impacts brought by these changes as well as their implications. It is natural to bring statistical forecasts and experts' contextual knowledge and experience together to increase forecast accuracy in order to reduce the risk of decision-making for tourism practitioners.

To establish a holistic analytical framework for integrating statistical forecasts with human judgment, both quantitative and qualitative analyses are applied. The quantitative analysis aims to examine the forecasting performance of statistical and judgmental forecasts from three dimensions: accuracy, bias, and efficiency. More specifically, the quantitative analysis presents the forecasts and analyses their forecasting performance by applying quantitative approaches such as statistical tests (parametric and nonparametric tests), correlation and regression analysis. The qualitative analysis, mainly through in-depth interviews, investigates the reasons for inaccuracy, bias, and inefficiency and explores the judgmental behaviour of tourism forecasters. The remainder of this chapter summarizes the major findings and implications from Chapter 5 and addresses the study's limitations as well as potential research directions for tourism researchers and industry practitioners.

6.2 Major Findings and Implications

6.2.1 Effectiveness of implementing judgmental adjustments

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 iteration. The results obtained by APE were consistent with those from MAPE and RMSPE, suggesting that these findings were not subject to error 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. The relative accuracy of the statistical forecasts and the judgmentally adjusted forecasts was also compared with the simple Naive forecasts. The Theil's U statistics for five of the six markets were below unity for the two rounds, suggesting that the unadjusted and adjusted forecasts were better than the Naive forecasts for these markets. The only case reporting a U statistic larger than one was the Mainland market. Although the U statistics reduced significantly after the experts' judgmental adjustments, they were still above unity for the Mainland forecasts.

The judgmental adjustments were more effective for short-haul markets as more remarkable improvements in accuracy were observed among the short-haul markets than among the long-haul markets. Specifically, the statistical forecasts achieved higher accuracy in the long-haul markets (Australia, the UK, and the USA) than in the short-haul markets (China, Japan, and Taiwan). A number of reasons could account for such a difference in accuracy improvement. First, the level of data variability can affect the effectiveness of experts' adjustments in making arrivals forecasts in Hong Kong. The judgmentally adjusted forecasts appeared to be more accurate than the statistical forecasts for arrival series with higher variability: the short-haul markets tended to be more volatile than the long-haul ones. This suggests that human interventions are likely to be more beneficial to accuracy improvement for arrival series with high volatility. For low-variability series, experts' judgmental adjustments would probably harm forecast accuracy if the initial statistical forecasts are 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.

One useful finding emerging from this study is that the forecasts adjusted by the experts were those most in need of adjustment. In the interviews, it was reported that the majority of the experts were inclined to adjust forecasts for markets that they had more market knowledge of and thus felt more confident in their adjustments.

The findings from the hypothesis tests also proved the value of the Delphi approach in enhancing the effectiveness of judgmental adjustments to statistical forecasts. Quantitative analyses such as statistical tests, regression analysis, and correlation analysis were employed to examine the values and changes of error measures and raw forecasts. Compared to the initial round, significant improvements were achieved after the experts' adjustments in the second round. The results of the parametric and nonparametric tests showed that significant improvements in accuracy were not only observed for the whole Delphi panel but also for individual panellists. Regression analysis, which can show the degree of association between the forecasts and actual values, was conducted to gain additional insights into the relative performance of the statistical and judgmentally adjusted forecasts. The regression results indicated that the final Delphi forecast was likely to be a better predictor of actual visitor arrivals in Hong Kong than the group forecasts in the initial round. Furthermore, the findings from the interviews demonstrated that the majority of the experts utilized other experts' comments to check whether they had overlooked some important factors that were not considered in their initial adjustments or whether they may have adjusted for the same reasons but with a different magnitude. The exchange of experts' views is likely to help Delphi participants produce more efficient forecasts.

It should be noted that the above findings were made based on the evaluation period 2011Q2–2012Q2. An examination of the degree of improvement over the accuracy criteria (APE, MAPE, and RMSPE) showed that as the forecasting horizon extended, the overall accuracy averaged from the six source markets tended to decrease. Furthermore, there was a decreasing trend of accuracy improvement over the evaluation period, indicating that the experts' judgmental forecasting ability reduced over time. In addition, for all of the forecasting horizons examined, the final

Delphi forecasts were more accurate than the initial consensus forecasts and statistical forecasts, suggesting that forecasting risk is likely to be reduced through the use of the structured group technique as this technique leads to more accurate forecasts.

According to the findings from the feedback survey and in-depth interviews, a number of factors account for the accuracy improvement in this study. First, the application of a Delphi procedure can structure and quantify experts' knowledge and experience into the forecasting process. Second, the use of an innovative Web-based forecasting system (HKTDFS) facilitates experts' judgmental adjustment process. Third, accurate econometric forecasts provide a solid foundation for experts' adjustments. Last but not least, a Delphi panel is composed of a number of experienced industry practitioners and academic researchers with high forecasting expertise.

6.2.2 Bias and inefficiency of judgmental adjustments

All forecasts are made under varying degrees of uncertainty, with no meaningful prediction ever being completely certain. 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 hypotheses *H2a* and *H2b* 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. This overforecasting tendency was observed in predicting visitors from five markets (Japan, Taiwan, Australia, the UK, and the USA) while the tendency for underforecasting was detected in estimating the volume of Mainland visitors.

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 interview findings showed that the Delphi technique helped the experts to make better use of the information available to them but could not change their use of heuristics in making adjustment decisions. The use of different heuristics can produce different biases, such as anchoring and recency. The regression analysis conducted by regressing the adjusted forecasts over actual arrivals indicated that the experts' judgmentally adjusted forecasts were highly anchored on the baseline (statistical) forecasts. The findings from the interviews provided further evidence that the experts had a high reliance on baseline forecasts. Most of the interviewees reported that their revised forecasts did not deviate too much from the initial statistical 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. Furthermore, the provision of a variety of forecasts made by different forecasting methods may also be helpful in reducing such bias.

Some of the experts stated that they made adjustments to incorporate the impacts of recent events and foreseeable events over the forecasting period. For example, the interview findings showed that the adjustment decisions of some of the experts were largely driven by assessing the possible impacts of specific events (e.g.

Japan earthquake, 2012 London Olympic Games) that could be easily recalled from memory. Consequently, these experts brought the availability bias into their forecasting adjustment process. This study also found that the experts still made adjustments even when they were provided with highly accurate forecasts. An additional analysis was conducted to test the relationship between the size of adjustments and accuracy improvement, but no clear association was identified.

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 the anchor development, selection of baseline forecasts, and provision of feedback. Since no studies testing the bias and efficiency 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.

Consistent with the prior literature, this study also found that judgmentally adjusted forecasts based on statistical forecasts are inefficient, suggesting that experts fail to incorporate all of the pertinent information from their own past forecasts and forecast errors. Given the inefficiency of judgmentally adjusted forecasts, future work should investigate how to improve the effectiveness of forecasts through the provision of guidance in the HKTDFS.

6.2.3 Conditions for using judgmental adjustments

If experts have knowledge about big recent changes, judgmental adjustments of the current status are likely to improve accuracy. Only when tourism forecasters have important information about the market that is not available in the statistical forecasts or the increase or decrease is a harbinger of a fundamental change should they use experts' judgment to make adjustments. However, when the statistical models are quite reliable and capable of producing highly accurate forecasts and experts do not have much extra model information to contribute, it is better to rely on statistical forecasts rather than human interventions.

In addition, the relative accuracy of judgmentally adjusted forecasts is likely to vary according to the level of data variability. The findings obtained from testing hypotheses *H5a* and *H5d* revealed that the judgmental adjustments of visitor arrival forecasts were more beneficial in predicting markets with high uncertainty and volatility than those with lower conditions; for example, significantly more gains in terms of error reduction through judgmental adjustments were obtained from predicting arrivals from the Mainland than the UK because the Mainland market is much more vibrant and volatile and thus it is much more difficult to predict the demand.

6.2.4 Exploring the underlying assumptions behind experts' adjustments

The findings from the in-depth interviews suggested that the experts had different views on what the underlying assumptions behind their judgmental adjustments were. In this study, the experts' assumptions were classified and summarized into four main categories: market knowledge and characteristics, tourist behaviour, statistical forecasts as the foundation, and feedback from other Delphi experts.

Market knowledge and characteristics was found to be the most important assumption determining the judgmental adjustments of the Delphi participants in this study. A number of factors regarding market environment were considered by the respondents: economic factors (e.g. economic growth, inflation, income levels, and employment), social factors (e.g. population), political factors, and the impacts of specific events in the near future. Among these factors, the experts regarded economic conditions in the origin source markets as the key factor guiding their judgmental forecasting decisions.

The second important assumption considered by the Delphi experts in making their adjustments was tourist behaviour, including the characteristics of tourists from individual source markets, travel patterns, type of tourists, and visiting purposes. It was also found that the experts generally anchored on baseline (statistical) forecasts and made revisions accordingly. Comments and views from other experts also played a role in the experts' adjustment activity. The exchange of expert opinions facilitated the information sharing process and thus improved the efficiency of forecasts.

6.2.5 Usefulness of applying the Delphi procedure

Focusing on the advantages of the Delphi method employed in the present study, the most obvious and convincing argument for using this method is its potential for eliciting and combining expert judgments to produce forecasts which are substantially more accurate than those of individual experts and traditional groups and somewhat more accurate than those of statistical groups (in which the judgments of noninteracting individuals are combined).

The empirical results of this study indicate that the Delphi procedure is likely to provide more accurate forecasts than those produced by a statistical group: the group forecasts in Round 2 were more accurate than those in Round 1. The statistical test results confirmed that forecast accuracy improved significantly over rounds irrespective of the error measures used. This result held true for both group consensus forecasts and individual experts' forecasts. The regression analysis further confirmed that the group forecasts in Round 2 were a better predictor of actual arrivals than the group forecasts in Round 1. The findings from the interviews showed that the use of the Delphi approach helped to assist with the experts' judgmental adjustments (e.g. information exchange and sharing among panellists regarding the rationales behind their adjustments; having the chance to revise their original forecasts based on new information and comments from other experts).

One problem associated with the Delphi technique is how to combine expert judgments. In the existing literature, both mean and median values are used as the consensus measure to aggregate group forecasts. This study compared the MAPE, RMSPE, and Theil's U statistics values of the consensus forecasts obtained from the median, the equally weighted mean, and the self-rated expertise weighted mean to test whether the findings differed. It was found that using the equally weighted mean as the consensus measure to obtain final Delphi forecasts led to more accurate forecasts than those produced using the median and the weighted mean by self-rating expertise. This finding indicates that the conventional mean measure can be used as a reasonable consensus estimate of the group judgment if accuracy is the top concern.

One key factor in the successful application of the Delphi technique is the selection of the panellists. Actually, a few of the industry experts interviewed expressed concern about the selection of Delphi participants for this study. The majority of the interviewees agreed that it is crucial to select experts with rich forecasting experience and pertinent market knowledge and information as qualified panellists; otherwise, the final Delphi forecasts may be diluted. An examination of the forecasting performance of individual experts showed that the accuracy of approximately half of the panellists was below the average level as suggested by MAPE and RMSPE. This finding confirm that to achieve higher forecast accuracy, it is fundamental to select a panel of qualified experts who can contribute to accuracy improvements. Moreover, the number of panellists may not be a critical factor for accuracy improvement in Delphi forecasting; rather, the quality of experts' judgments is much more valuable than the quantity.

A comparison of forecasting performance between the industry experts and the academic experts showed that the industry forecasters provided significantly more accurate forecasts than the academic group irrespective of the error measures used. Hence, it is desirable to include more industry practitioners than academic researchers in a Delphi panel for future tourism demand forecasting in Hong Kong.

Identifying experts on a basis of self-rated expertise may not provide better estimates than simply taking the average of individual experts' judgments. A subgroup analysis based on self-rated forecasting expertise found that while the experts tended to make more accurate demand forecasts than the less-expert subgroup, these differences did not reach statistical significance. This indicates that a simple external measure of expertise can be developed which would select experts with higher self-rated expertise who could probably provide better adjustments than those with lower self-rating scores. It may be argued that the self-rating scale used in this study is not a measure of expertise in a specific forecasting task (i.e. forecasting visitor arrivals from specific source markets in Hong Kong) but rather a generalized construct of expertise in tourism demand forecasting.

6.2.6 Use of a forecasting support system

The use of human judgment in forecasting can suffer from people's limited capability to deal with complex problems and large amounts of information. It is thus helpful to use a procedure or decision support system to structure human judgment in both decision-making and forecasting processes. This study utilized a tourism demand forecasting system (i.e. HKTDFS) to structure the judgmental adjustment process through an online Delphi procedure. This allowed the Delphi participants in this study to generate independent predictions into the system. Such a design feature might be helpful to mitigate the pressure towards bias. Although following the Delphi procedure is time consuming, the use of such a structured process helps to produce more accurate forecasts.

One aspect of the structuring process that can affect forecast accuracy is the data presentation format. There is a mixture of evidence that can be used to assess the impact of graphs and tables on forecast accuracy. This study did not directly evaluate the effects of graphical versus tabular data presentation on forecasting performance but rather considered the effects of using both formats. The feedback surveys undertaken by the student participants in the pilot survey and the respondents from the in-depth interviews both reported the positive influence of graphical and tabular formats (particularly graphical displays) in helping them to understand the characteristics of a time series and facilitating their judgmental adjustment process. The majority of the experts interviewed preferred to have both tables and graphs, although graphs appeared to be more intuitive and easier to understand and operate.

The majority of the interview respondents indicated that the information guidance in the HKTDFS, such as graphical and tabular interfaces, the provision of recent observations and statistical forecasts as reference values for adjustments, the provision of information regarding forthcoming events over the forecast period, and feedback and comments shared by other experts, helped them most with their adjustments. The findings obtained from the interviews suggested the need to strengthen the relationship between tourism managers and researchers in order to jointly develop a forecasting support system that would encourage the managers to engage in forecasting at a deeper level. The results of the current study should encourage the successful implementation of effective forecasting processes for public and private tourism organizations/companies.

6.2.7 Contributions of the study: From the experts' perspective

Given the importance of incorporating judgment into their forecasts to decision makers and forecasters in the tourism industry and the importance of making accurate and unbiased forecasts to businesses and society, this study makes valuable contributions to the existing tourism demand forecasting literature by examining tourism forecasters' forecasting performance and adjustment behaviour through quantitative and qualitative approaches.

The empirical results from the pilot and main Delphi surveys both confirmed the value of integrating judgments into statistical forecasts in tourism demand forecasting. The experts interviewed all agreed that it is necessary to integrate experts' judgment into statistical forecasting methods given the relative strengths and weaknesses of each method. This integration should be particularly beneficial in tourism demand forecasting as the tourism industry is facing full dynamics and high uncertainty. The main benefit of the integration is that it improves forecast accuracy because the integration of the two forecasting methods enables the incorporation of up-to-date information into the forecasting process and the detection of changing patterns in a time series.

According to the experts' views, this study makes three major contributions to the Hong Kong tourism industry. First, it has provided reliable and accurate visitor arrivals forecasts by source markets for Hong Kong, and these forecasts offer a scientific foundation for industry practitioners' strategic decision-making. Second, it is the first forecasting exercise in the tourism forecasting context that has attempted to integrate tourism forecasters' judgmental predictions with statistical forecasts produced by advanced econometric forecasting models. Third, it provides useful insights and valuable suggestions/recommendations for tourism researchers and practitioners on establishing a more effective and user-friendly tourism demand forecasting support system in Hong Kong. Lastly, 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.

6.3 Study's Limitations and Potential Research Directions

As with any research, this study has certain limitations that need to be noted and addressed in future research where possible. The empirical setting of the present study was specific in order to forecast visitor arrivals in Hong Kong and therefore it may not be strictly comparable with the settings of earlier efforts using other types of data series in other fields. The findings from this study not only add to our current knowledge on the subjects of judgmental adjustments to statistical forecasts and tourism forecasters' judgmental behaviour but also introduce promising future research directions in relation to integrating human judgment into statistical forecasting methods in the tourism context.

Although one should proceed with caution in attempting to generalize the findings from this study to other types of tourism forecasting tasks, this study provides a basis for extending future research on forecasting tourism demand for other destinations because the research hypotheses of this study are well-established hypotheses according to the relevant studies in the existing literature. Therefore, the findings obtained should not be regarded as mere statistical coincidences. Furthermore, the findings from this study could be comparable with prior studies concerned with judgmental adjustment based on statistical forecasts. However, there is always the chance that other studies conducted under different research backgrounds and business settings will not produce findings that are the same as, or similar to, those reached by this study.

The remainder of this section is divided into four subsections. The first three subsections discuss three major limitations of the current study: the sample size issue, the variable selection issue, and the issue of supply constraints in tourism demand forecasting. The last section addresses future work to further improve the existing HKTDFS.

6.3.1 Sample size issue

Unlike in experimental studies, one could not ask Delphi participants to provide predictions for historical data. Instead, panellists need to provide genuine arrival forecasts. This will inevitably limit the sample size used for forecasting evaluation. Given the project timeline, the forecasting horizon for evaluation purposes was relatively short in this study as the study could only cover the time span of 2011Q2−2012Q2 (five quarters) for the six source markets with a total of 30 observations.

Although a longer span of data would have been desirable, the maximum span available at the time at which all of the forecasts were made (i.e. June 2012) was five quarters. Consequently, the implementation of a more comprehensive statistical analysis was limited in some circumstances due to the small sample size. However, this study attempted to increase the sample size by making use of the adjustment data from individual Delphi experts, and this helped to extend the depth of the quantitative analysis. The examination of forecasting performance not only utilized forecasts from consensus groups (i.e. group forecasts in Rounds 1 and 2) but also distilled information from the adjustment data aggregated by different categories, such as such as source markets, forecasting horizons, and expert groups.

While sample size is important, its value can be overstated as sampling error is only part of the total forecast error (Armstrong, 2001b). However, the sampling error is important when small samples are used. As not much is known about the sampling distribution of actual postsample forecast errors, the generalizability of the results of this study may suffer.

6.3.2 Issue of the variable selection

A second limitation is that some variables (both dependent and independent variables) remained outside the scope of this study. Due to the data availability and time constraints of the project, this study only focused on forecasting visitor arrivals from six source markets in Hong Kong. One of the main reasons for choosing only one forecasting variable, namely visitor arrivals, was the experts' familiarity with the forecasting tasks associated with this tourism demand measure. The feedback survey in Section 4.4 of Chapter 4 reported that two thirds of the student participants took around 20 to 40 minutes to complete the survey. The interview findings also demonstrated that the Delphi participants thought that it was appropriate to set up forecasting tasks with six markets. Yet, a comprehensive Delphi survey covering more tourism demand measures, such as tourism expenditure and hotel room nights, over a longer evaluation period would require more effort and a longer time period, which would degrade the accuracy of responses because of respondent fatigue. Hence, the Delphi survey in this study was designed to limit forecasts to six source markets.

As shown in Chapter 3, the key determinants used to model and forecast visitor arrivals in Hong Kong are income level, own price, substitute price variables, the lag of visitor arrivals, and dummy variables. Some other explanatory variables, such as marketing expenditure and fuel price variables, were not considered in this study because it is difficult to obtain quarterly data on these variables.

6.3.3 Issue of the supply constraints

In the existing tourism forecasting literature, the centre of attention is the demand side and the influencing factors of tourism demand, but future research should incorporate factors from the supply side into demand forecasting as well. According to the interview findings, the respondents expressed a common concern regarding the issue of the supply constraints on tourism demand forecasting. Some of them questioned the projected growth of arrivals forecasts and its sustainability as the statistical forecasts were too optimistic and robust and apparently too promising as such growths cannot be nonstop. They further argued that the supply constraints, which are probably in the form of shortages of hotel accommodation and limited passenger transportation capacity and flight capacity, should be included when making arrivals forecasts in Hong Kong. Although the inclusion of the lagged dependent variable into the ARDL-ECM has been justified as one way to accommodate supply constraints, it is worth making further efforts to include supply constraints in the existing forecasting models and this is probably a good starting point for making sound tourism forecasts.

6.3.4 Future work in further developing the HKTDFS

The empirical results of this study showed that the HKTDFS is a user-friendly system and is capable of providing reliable and accurate forecasts. As indicated by the interview respondents, there is a need for future research into developing a more effective forecasting support system in Hong Kong's tourism industry. Suggestions on further ways to improve the effectiveness and functional ability of the HKTDFS are elaborated in the remainder of this section.

(1) Choice of forecasting methods

The discussions below about the choice of forecasting methods is divided into three parts: econometric models to make baseline forecasts, benchmark models for evaluation, and combined forecasting.

To increase forecast accuracy, it is worth attempting to make forecasts using more advanced econometric forecasting techniques such as the TVP model and its variants. Unlike the ARDL approach, the TVP approach relaxes the assumption of parameter constancy, and the behavioural change of tourists over time is traced using a statistical estimator known as a Kalman filter. The appropriateness of the TVP approach in tourism demand modelling has been recorded in the existing tourism forecasting literature. For example, Li, Song, and Witt (2006) developed a time-varying parameter linear almost ideal demand system (TVP-LAIDS) model and concluded that such a model improves forecasting performance remarkably compared to its original static version of fixed-parameter error correction counterparts in terms of modelling and forecasting the demand for tourism in Western European destinations from UK residents.

The performance of different forecasting methods is likely to vary with the type of data, the forecasting horizon, and the error measure applied. The findings from the interviews in this study suggested that in addition to accuracy, ease of use and implementation is another important factor considered by tourism forecasters when selecting a forecasting method. The industry practitioners interviewed in this study preferred to use simple methods due to time concerns and a lack of understanding of more sophisticated forecasting methods. Hence, the provision of various options, including both simple and sophisticated forecasting methods, may attract tourism practitioners to use a forecasting system (HKTDFS) and get more involved in the forecasting process. Based on the above considerations, a group of benchmark models, such as Naive models, exponential smoothing models, Box-Jenkins time-series models (e.g. ARIMA), and simple regression, should be featured within the HKTDFS in order to increase the flexibility of options for making baseline forecasts and for comparison purposes. If the selected forecasting methods are found to consistently produce inaccurate final forecasts, these methods should not be used for future forecasting.

Shen, Li, and Song (2008) found that combined forecasts overall played an important role in improving forecast accuracy over different forecasting horizons. More recently, Shen, Li, and Song (2011) concluded that combined forecasts generally outperform the best individual forecasts and the performance of combined forecasts is associated with the performance consistency of their constituent forecasts. Similarly, Song et al. (2009) found that combined forecasts are significantly more accurate than the average single-model forecasts across all forecasting horizons examined. They also suggested that the combination could be more beneficial in longer-term forecasting. Slightly different from the above findings, Wong et al. (2007) showed that combined forecasts do not always outperform the best single-model forecasts, but do at least outperform the worst single-model forecasts. They also illustrated that the relative performance of the combined forecasts varies according to the origin-destination under consideration. The forecasting combination literature in tourism has suggested that combined forecasting approach can generate more accurate baseline forecasts than any single model.

As accuracy is the most important criterion in selecting a forecasting technique, built-in measures that allow for comparing the accuracy of baseline forecasts against benchmark forecasts should be enabled in the HKTDFS. In addition to accuracy, the HKTDFS should also be able to include indicators showing whether or not judgmental forecasts are biased and inefficient; this would enable forecasters to obtain information about their past forecasting performance which might help to improve their future forecasting.

Furthermore, the interview participants also suggested that if the HKTDFS decided to provide forecasts made by different forecasting methods, a brief explanation of each forecasting model and an indication of the circumstances where its application is appropriate should also be provided. Alternatively, some experts stated that to better facilitate experts' interaction with a forecasting system, it would be desirable if the system could provide its users with an option to automatically select a suitable forecasting model that gives the best goodness of fit or the least forecast error, as judged from past data, and generates forecasts based on that method.

The HKTDFS could also provide its users with an option to combine any two or more of the above quantitative models. The rationale for this is that combining two or more quantitative methods normally leads to an improvement in forecast accuracy. The advantage of combined forecasting is not that the best possible combinations outperform the best individual component forecasts but that in terms of reducing forecasting risks in practice, it is better to combine forecasts than to select a single forecasting model. For many practitioners, a realistic option that provides better results than judgmental adjustment is the mechanical integration of judgmental and statistical forecasts.

(2) Design and implementation of a more effective guidance system

Although the participants from the pilot and main Delphi surveys suggested that the current HKTDFS is a user-friendly system, there is still room for further improvement in the future. As already presented in Chapter 3, the design of the *Forecasting Adjustment* module is based on the two concepts of *restrictiveness* and *guidance*. For consistency, the recommendations made by the Delphi participants to improve the effectiveness of the HKTDFS are summarized and analysed according to the two broad categories above.

The experts interviewed made more suggestions for relaxing the restrictiveness of the HKTDFS than on how to improve the guidance given in the system to assist judgmental forecasting adjustments. With regard to restrictiveness, comments were made on four aspects of the system: data and variables, adjustment options, decomposition, and easy-to-use design features. For example, a few experts suggested that making adjustments could be made more flexible and interactive by functions that enable drag and drop movements onto graphs or the revisions of numbers directly on the tables. Another expert thought that the system could add a function to force a participant to provide justifications when there is a huge difference between the participant's adjustment and the baseline (statistical) forecasts.

The interview findings with respect to guidance strategies have inspired the system designer to build a more supportive and deliberate tourism forecasting system by adding more decisional guidance into the HKTDFS. As indicated by the interview findings, decisional guidance can be provided by employing three design strategies: informative, suggestive, and predefined guidance. Such a guidance system will help users to make better forecasting decisions. Furthermore, it will provide the ability to monitor experts' reason feedback over time to evaluate users' forecasting performance and explore their adjustment behaviour. In short, a built-in guidance system can be designed and implemented to further enhance the user-friendliness and effectiveness of the HKTDFS.

(3) Documenting reasons for feedback

Due to the restriction of the project's timeline, this study only focused on evaluating a relatively short period of the forecasting performance of experts' judgmental adjustments and econometric forecasts. It is recommended that experts should be invited to make their judgmental adjustments periodically so that their forecasting performance can be monitored over time. Specifically, results such as what types of adjustments lead to greater improvements in accuracy over time and what types of adjustments are more effective can be evaluated. As a result, the HKTDFS will be able to trace users' actions when using the system in order to accumulate knowledge concerning users' forecasting behaviour.

A feature that allows users to document their forecasting activities could serve as a useful resource for selecting qualified experts to conduct judgmental forecasting tasks in the HKTDFS. Therefore, it is recommended that a new feature that enables forecasters to keep records of all the adjustments they have made and the reasons for making them be built into the system. Further empirical evidence will be required thereafter to investigate the relationship between the documentation process and forecast accuracy, in order to examine the validity and effectiveness of building this new feature.

(4) Provision of an online communication forum

The findings from the in-depth interviews indicate a need for establishing interactive communication among tourism forecasters. The inclusion of an online forum connected to the *Forecasting Adjustment* module may help experts to improve their forecasting performance, as it would provide rapid feedback from tourism forecasters who use the HKTDFS. An online forum can serve at least three information-sharing purposes. First, common driving factors or assumptions behind experts' predictions can be clearly identified and timely shared among a panel during a Delphi survey period, which would not only help all expert panellists to gain a better understanding about what they are doing but also aid them in better utilizing experts' experience and knowledge and updated information. Second, it would also help to provide a historical record of judgmental inputs for future evaluation. Third, the use of an on-line forum could possible reduce the bias brought by human judgment; for example, the forum could be used to reduce the bias of selective perception by asking experts with different backgrounds and experience to provide a holistic list of forthcoming events in the forecasting period via the forum discussions. Doing so could allow experts to avoid or reduce the negative impacts of underestimating uncertainty. Another benefit of an online forum would be that in addition to its anonymity, it would provide a more cost-efficient means of communication among Delphi participants. This enhanced communication might also help experts to focus more on the tasks at hand, ensure equal participation and stimulate the generation of ideas.

Appendices

Appendix A: Instructions for Delphi Survey (Round 1)

Introduction to the Delphi Survey

Experts are invited to make their adjustments to the quarterly statistical forecasts of visitor arrivals from three short-haul markets (i.e. **China, Taiwan, and Japan**) and three long-haul markets (i.e. **the USA, the UK, and Australia**) of Hong Kong over the period of **2011Q2 to 2015Q4.**

Events that need to be considered over the forecasting period include:

(1) Japan earthquake in 2011

The 2011 Tōhoku earthquake, also known as the Great East Japan Earthquake, was a magnitude 9.0 undersea mega thrust earthquake off the coast of Japan that occurred at 14:46 on Friday, 11 March 2011. It was the most powerful known earthquake to have hit Japan, and one of the five most powerful earthquakes in the world overall since modern record-keeping began in 1900.

(2) High-speed railway (January 2010-2015)

The 26-km long Hong Kong Section of the Guangzhou-Shenzhen-Hong Kong Express Rail Link runs from West Kowloon in Hong Kong to the boundary of Hong Kong and Shenzhen. The Express Rail Link will connect with the 16,000-km National High-speed Railway Network and will enhance Hong Kong's role as the southern gateway to the Mainland.

(3) 2012 London Olympic Games (27 July to 12 August 2012)

London will become the first city to officially host the modern Olympic Games three times, having previously done so in 1908 and 1948.

(4) Three New Themed Lands in the Hong Kong Disneyland to Be Introduced

November 2011	2012	2013
The Toy Story Land (It has been rumoured that the popular Toy Story Midway Mania attraction will be included.)	**Grizzly Gulch**	**Mystic Point**, is a new themed land in Hong Kong Disneyland.

(5) Upcoming Events & Festivals released by the Hong Kong Tourism Board

Please visit

http://partnernet.hktb.com/pnweb/jsp/comm/index.jsp?charset=en&pageContent=%2Fjsp%2Fdest%2Fef.jsp.

Note: The above events have not been taken into account in the statistical forecasts.

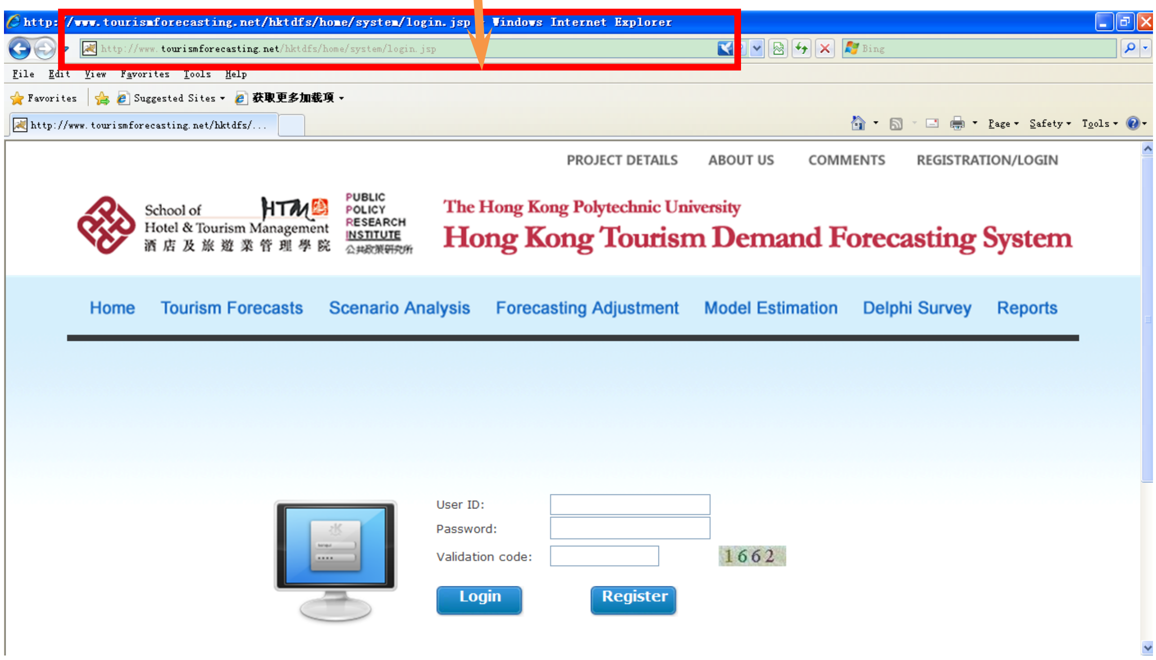
Guides to Making Forecasting Adjustments

The system is user-friendly, however, to ensure you can correctly use the system to make adjustments, the step-by-step instructions are provided as below.

Section 1: Access to the System

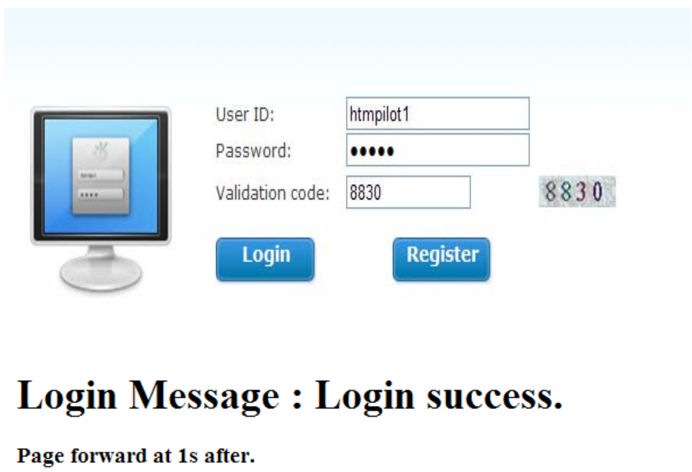
Please kindly follow the instructions below when you access the forecasting website to carry out the forecasting adjustments:

Input http://www.tourismforecasting.net/hktdfs/home/system/login.jsp in the address bar of the Internet Explorer on your computer. Please use Internet **Explorer 6.0** or above to visit this Web site.



Step 1: Input the **User ID** and **Password** in the pop-up window (please note that the username and password here are **case-sensitive**).

Step 2: Once you have logged onto the system, you will find the message appearing in the pop-up window: ***Login message: Login success***. Before the survey, you will be able to visit the Web site to get a feel for what the system is about. After you have gone through the website, you can carry out the forecasting adjustments as a Delphi expert.

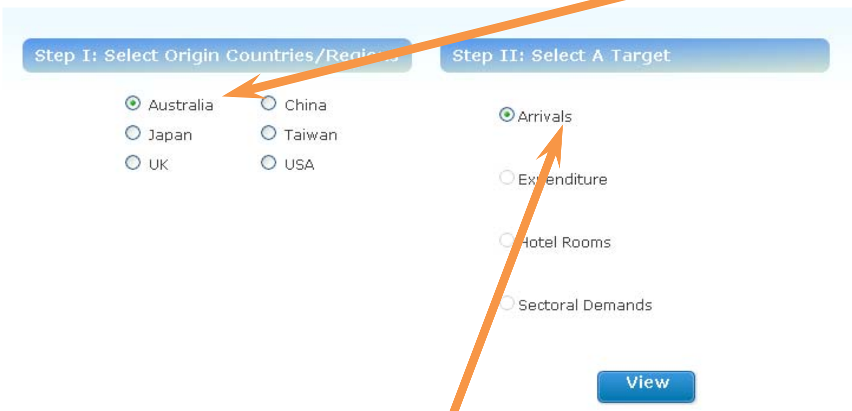


Section 2: Methods of Forecasting Adjustments

Please follow the instructions below to make your adjustments to the forecasts:

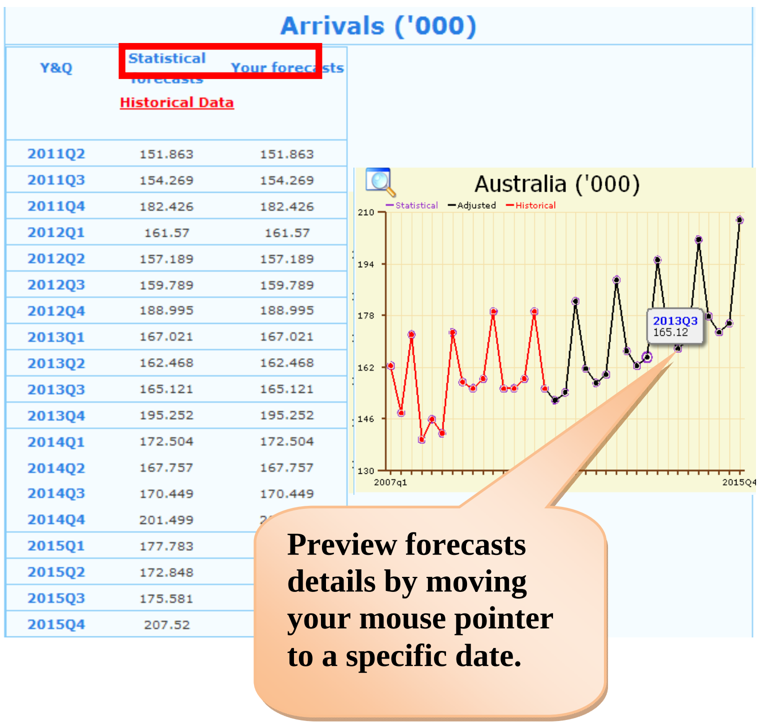
Step 3: Click "Forecasting Adjustment" on the main menu. Several options will be displayed under the "Forecasting Adjustment" tab.

Step 4: Please select one source market per time, for example, **Australia**.



Step 5: Please select the type of forecasts, "**Arrivals**", and click on "**View**" to continue.

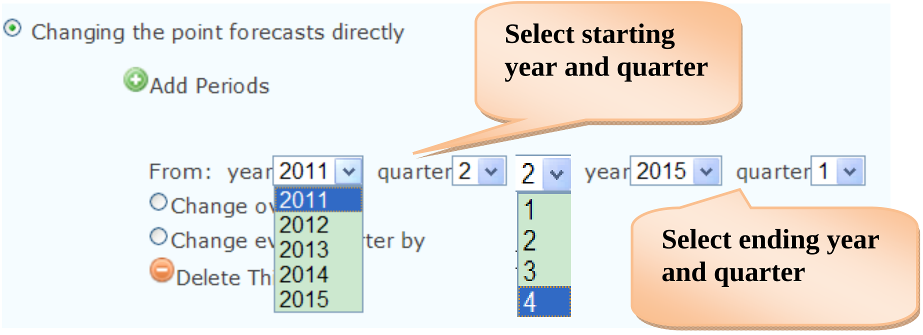
Step 6: By clicking on "**Historical Data**", you could see or hide the actual arrivals over 2007Q2-2011Q1. The column of "**Statistical forecasts**" is provided as the baseline forecasts. After making your adjustments, the adjusted forecasts will be presented in the column of "**Your forecasts**".


Arrivals ('000)

Y&Q	Statistical forecasts	Your forecasts
	Historical Data	
2011Q2	151.863	151.863
2011Q3	154.269	154.269
2011Q4	182.426	182.426
2012Q1	161.57	161.57
2012Q2	157.189	157.189
2012Q3	159.789	159.789
2012Q4	188.995	188.995
2013Q1	167.021	167.021
2013Q2	162.468	162.468
2013Q3	165.121	165.121
2013Q4	195.252	195.252
2014Q1	172.504	172.504
2014Q2	167.757	167.757
2014Q3	170.449	170.449
2014Q4	201.499	[illegible]
2015Q1	177.783	[illegible]
2015Q2	172.848	[illegible]
2015Q3	175.581	[illegible]
2015Q4	207.52	[illegible]

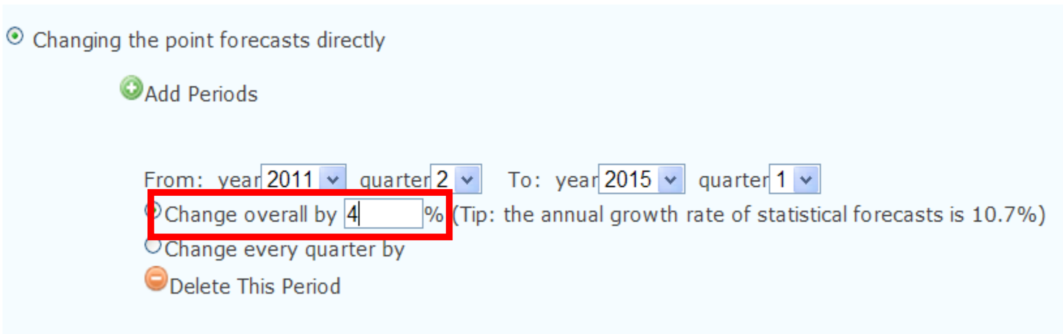
<u>Step 7:</u> Before you SUBMIT any adjustment, the Web page will only present the statistical forecasts in both tabular and graphical forms.

<u>Step 8:</u> Click on **Changing the point forecasts directly** and specify the forecasting period(s) for which you want to adjust the forecasts. To specify different forecast periods, you could click **Add Periods**.

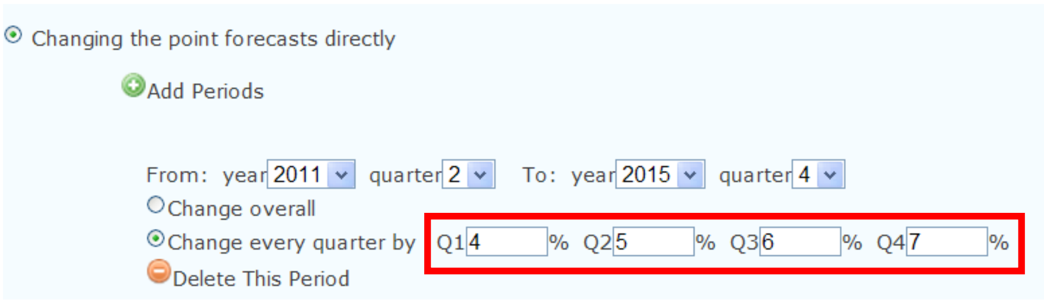


<u>Step 9:</u> Two options are available for you to make adjustments. **A percentage increase or decrease to the statistical forecasts for the period selected are required.** For example, the statistical forecast available in the system is 100, if you input 4%, the adjusted forecast will be increased to 104. If you input -4%, the adjusted forecast will be decreased to 96.

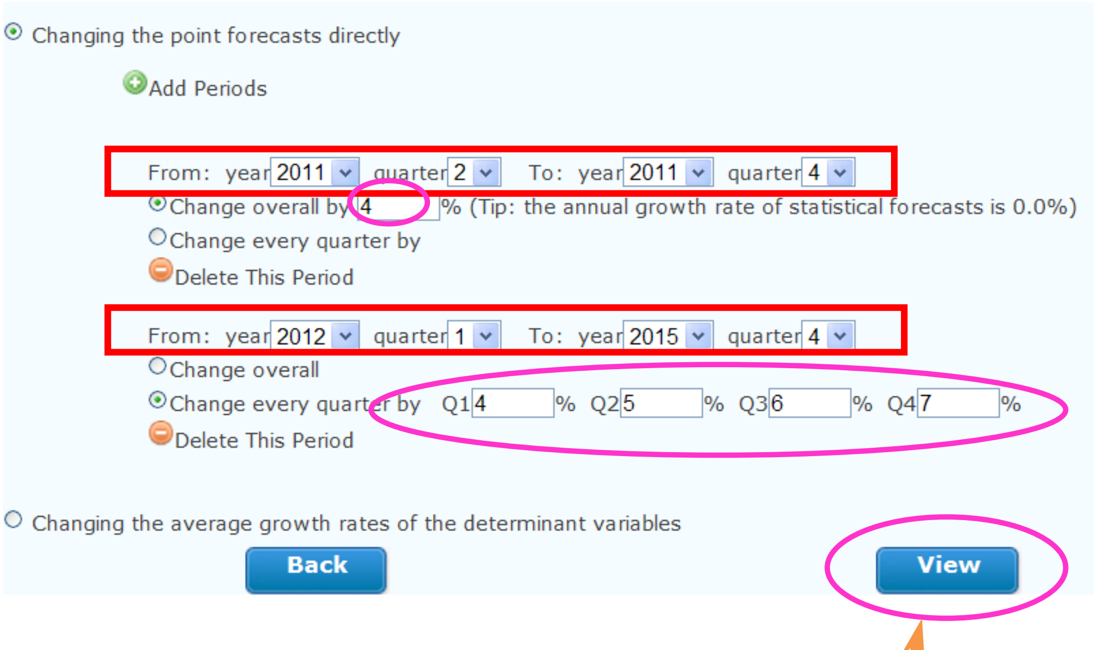
a) You could change the overall growth rates, for example, 4% over 2011Q2-2015Q4.



b) You could also change the growth rates by individual quarters, see the example as:

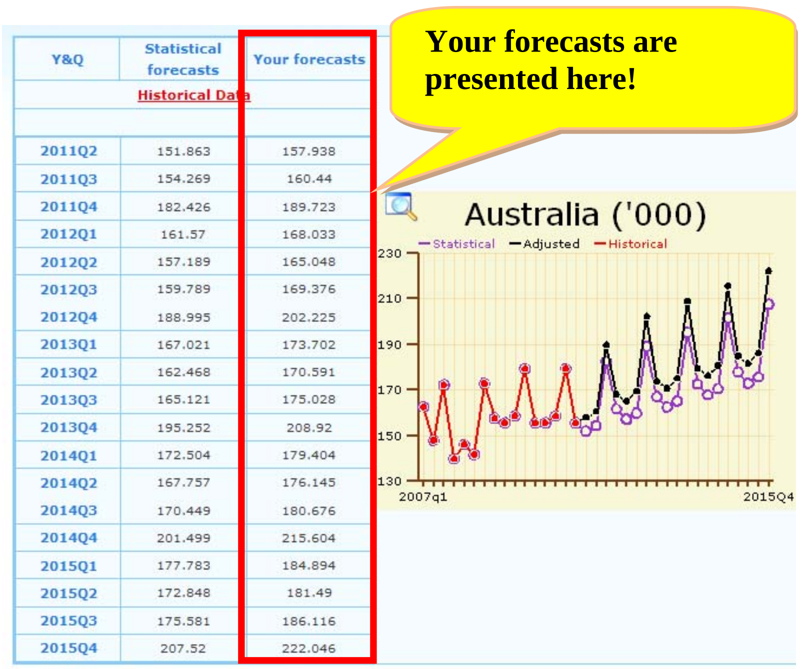


c) You could make use of two methods at the same time, for example, input 4% as an adjusted overall growth rate for 2011Q2-2011Q4; and 4%, 5%, 6% and 7%, for Q1, Q2, Q3, and Q4, respectively, over the period of 2012Q1-2015Q4.



Step 10: After inputting an adjustment to the relevant cell, click the "**View**" button to activate the adjustment.

Step 11: The adjusted results will be shown in the column of "**Your forecasts**" (or the ***black*** line in the graph) together with the **Statistical forecasts** (in **purple**) on the same page.



Y&Q	Statistical forecasts	Your forecasts
Historical Data		
2011Q2	151.863	157.938
2011Q3	154.269	160.44
2011Q4	182.426	189.723
2012Q1	161.57	168.033
2012Q2	157.189	165.048
2012Q3	159.789	169.376
2012Q4	188.995	202.225
2013Q1	167.021	173.702
2013Q2	162.468	170.591
2013Q3	165.121	175.028
2013Q4	195.252	208.92
2014Q1	172.504	179.404
2014Q2	167.757	176.145
2014Q3	170.449	180.676
2014Q4	201.499	215.604
2015Q1	177.783	184.894
2015Q2	172.848	181.49
2015Q3	175.581	186.116
2015Q4	207.52	222.046

Step 12: Please note that you need to input the reasons for adjustment in the textbox before you can submit it. Avoid using any operation symbols (e.g. +, -, *, /, ^) to start the paragraph. To click on "**Reset**", it will erase all inputs and you can insert new statements. If you want to revise your forecasts for any reason, you can click on the "**Back**" button to make another adjustment until you obtain the most satisfactory result.

Step 13: **IMPORTANT!!!** After making an adjustment, you need to SAVE the adjustment by clicking on the "**Submit**" button. If the following window pops up, click on the "**OK**" button. This will complete the submission process.

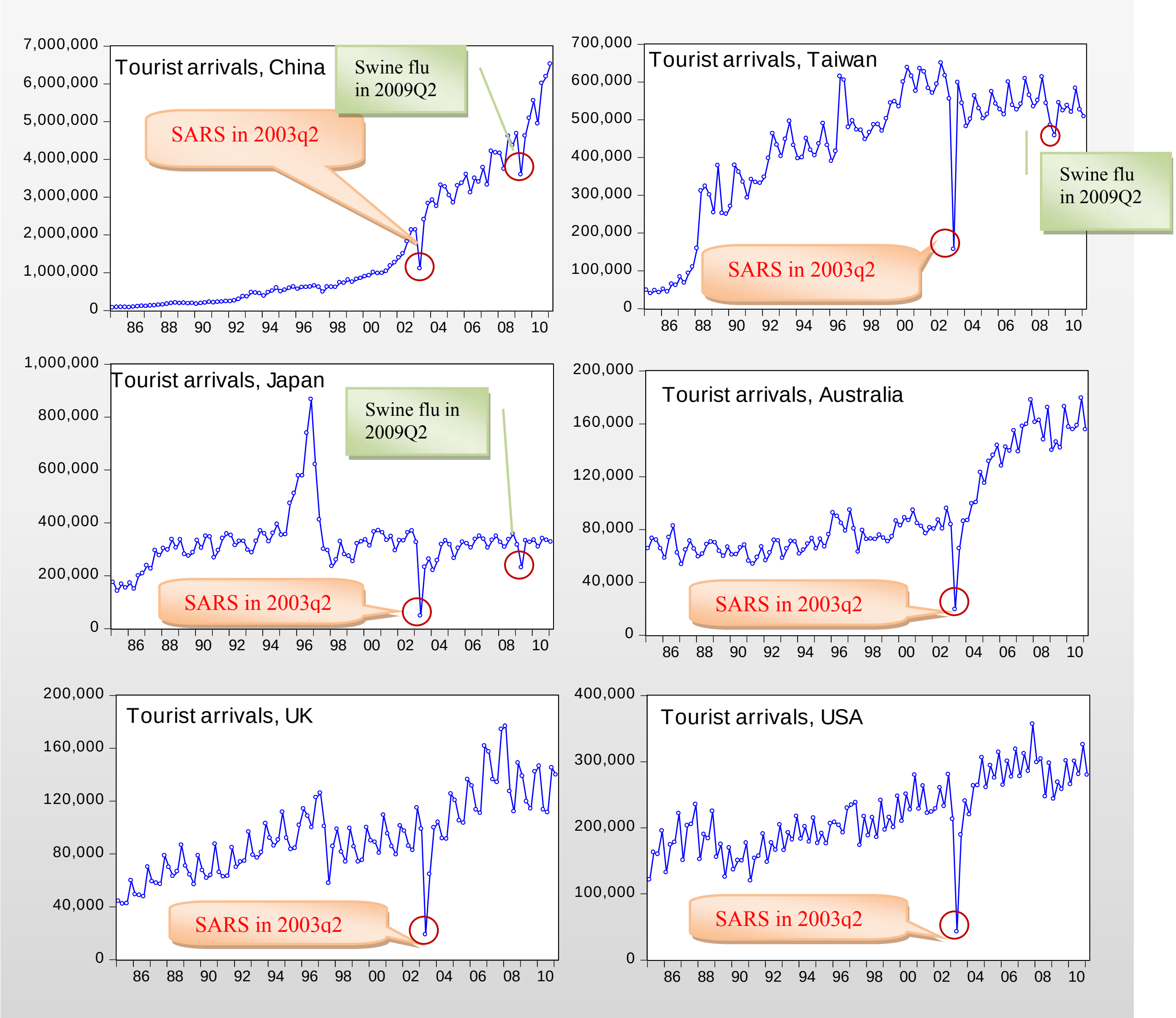
Step 14: After you have submitted your adjustments, you can view and change the adjustments anytime later during the same round of the Delphi survey. Please note that after the deadline for each round of the Delphi forecasts has passed, the submitted results will be regarded as your final adjustments. Your forecasts in the current round, together with the summarized group forecasts, will be presented for your reference in the next or second round of the survey.

Step 16: To move on to a new source market, please click on "**Forecasting Adjustment**" in the main menu. Select "**China**" as origin country and "**Arrivals**" as the type of forecasts, then click on "**View**" to continue.

Repeat the same procedure (from Steps 1-15) when making the adjustments for the other four source markets (i.e. Taiwan, Japan, USA, and the UK).

You can work on individual source market per time but remember always **SUBMIT** your job! For your reference, the annual and quarterly growth rates of visitor arrivals are attached in the following table.

Appendix I: Historical trends of visitor arrivals (1985Q1-2011Q1)



Appendix II: Projections for the real GDP growth rates and exchange rates

Categories	Geographies	2010	2011	2012	2013	2014	2015
Real GDP Growth - % growth	Australia	2.7	*3.0*	*3.5*	*3.5*	*3.3*	*3.2*
Real GDP Growth - % growth	China	10.3	*9.6*	*9.5*	*9.5*	*9.5*	*9.5*
Real GDP Growth - % growth	Japan	4.0	*1.4*	*2.1*	*1.7*	*1.5*	*1.3*
Real GDP Growth - % growth	Taiwan	11.0	*5.4*	*5.2*	*5.1*	*5.0*	*4.9*
Real GDP Growth - % growth	UK	1.3	*1.7*	*2.3*	*2.5*	*2.5*	*2.6*
Real GDP Growth - % growth	USA	2.8	*2.8*	*2.9*	*2.7*	*2.7*	*2.7*
Exchange Rates -A$ per US$	Australia	1.1	*1.0*	*1.0*	*1.0*	*1.1*	*1.1*
Exchange Rates - RMB per US$	China	6.8	*7.0*	*7.0*	*7.0*	*7.1*	*7.1*
Exchange Rates - ¥ per US$	Japan	87.8	*82.3*	*82.5*	*81.9*	*81.0*	*80.1*
Exchange Rates - NT$ per US$	Taiwan	31.6	*29.3*	*28.6*	*27.9*	*27.2*	*26.6*
Exchange Rates - £ per US$	UK	0.6	*0.6*	*0.6*	*0.6*	*0.6*	*0.6*
Exchange Rates - US$ per US$	USA	1.0	*1.0*	*1.0*	*1.0*	*1.0*	*1.0*

Source: International Monetary Fund (IMF).
Note: Figures in italic are forecasts.

-- End --

Appendix B: In-depth Interview Guide

Interview Date: ____________________, 2012
Time: ____________________
Respondent: ____________________
Organization: ____________________

Section 1: Introduction (3 minutes)

I want to thank you for taking the time to meet me today. My name is Vera and I would like to talk to you about your experience participating in my PhD project. Our purpose today is twofold:

- Investigate reasons why our forecasting approach can produce accurate forecasts;
- Provide suggestions to improve the forecasting performance of the proposed integration framework in tourism demand forecasting.

The interview will take around 30 to 45 minutes. I will be taping the session because I do not want to miss any of your comments. Although I will be taking some notes during the session, I cannot possibly write fast enough to get it all down. Because we are on tape, please be sure to speak up so that I do not miss your comments.
Feel free to make any negative or positive comments about any of the things we will be discussing today. This is a free flowing discussion and there are no right or wrong answers. Everything that you say here will be kept strictly confidential. This means that your interview responses will only be used for my PhD project and I will ensure that any information I include in my thesis does not identify you as the respondent.
Remember, you do not have to talk about anything you do not want to and you may end the interview at any time. Are there any questions about what I have just explained?
Note: Bring with the accuracy report, summarized reports (1st and 2nd rounds) for Delphi survey, and instructions.

Section 2: A review of judgmental forecasting in the HKTDFS (2 minutes)

- Display some screen shots of the system and briefly explain the forecasting procedure. If Internet is accessible, conduct an on-line demo.
- Present the historic trends of six source markets to the interviewee.

Section 3: Key questions (30 minutes)

Section 3A: Forecasting practice in your own organization

(1) [For industry experts] Does your organization use forecasts based on statistical methods to predict the demand?

Probing question: Are these forecasts made by your organization or by external organizations?

- If they make their own forecasts, then ask: What forecasting methods does your organization use (if any)?
- If not, what forecasting method(s) did the external organizations use?

(2) [For industry experts] Have you conducted forecasting for your organization?

Probing question: Which forecasting method did you usually use?

[For academic experts] Have you conducted any research based on statistical methods to predict the demand?

Probing question: What forecasting methods did you use (if any)?

(3) What criteria do you use (or which factors) do you consider important when selecting a forecasting method and why?

(4) Have you checked the forecasting performance of the forecasts you generated (or used)? (Have you checked the forecasting performance of the forecasts your organization used?) [Why or why not?]

a. If yes: how did you evaluate the forecasting performance?

b. If no: do you think it is necessary to check the forecasting performance of the forecasts you generated (or used)?

Section 3B: Opinions about different forecasting methods

(5) What type of forecasts did you use more often, experts' predictions or statistical forecasts? [Please explain.]

(6) In your opinion, which method do you think is generally more accurate in tourism forecasting? [Please explain why.]

(7) Do you think it is necessary to combine these two types of forecasts? Please elaborate.

- Will you expect a combination of these two methods can produce more accurate forecasts?

Section 3C: Improve forecasting performance of the HKTDFS

(8) What were the underlying assumptions of your adjustments in the Delphi survey?

Probing question: (1) Are those assumptions different for individual markets? (2) Did you change your assumptions over rounds?

(9) [Display historic trends of six source markets to the respondent.] Among six source markets, which market(s) you thought were more difficult to predict compare to other markets? Why?

(10) In our system, we provide two adjustment options, one is to change the overall growth rates, and the other is to change the growth rates by individual quarters. When you made your adjustments, which option did you use more often?

Probing question: Do you think it is sufficient to provide these two options? Why/ why not?

(11) When you made your adjustments, did you check the historical data of visitor arrivals provided by the forecasting system? Why or why not?

- If yes, did you check the tables (on the left hand side) or graphs (on the right hand side)?

(12) Which display format did you use (tables, graphs, or both) when checking the forecasts from the forecasting system?

(13) Did you make your adjustment based on historical data or the statistical forecasts of the system, or both?

- If more on historic data: did you check the recent one-year data or the historical trend?

(14) What helped you the most when using HKTFDS to make your forecasts?

(15) What will you expect from the forecasting system in terms of helping you with the forecasting adjustments? [Any recommendations you can provide that would probably help you to make your adjustments? Would you give me an example?]

(16) What impact, if any, do you think this project had on the industry in which you work? [What contributions do you think this project had on the industry in which you work?]

(17) What recommendations can you provide for further improving HKTDFS?

Section 4: Wrap up (5 minutes)

We have covered a lot of ground in our discussion today.

(1) Is there anything else you would like to add?

I will be analysing the information you and others gave me and compiling a draft report for all interviews in two months. I will be happy to send you a copy to review at that time, if you are interested. Thank you very much for your time!

Appendix C: The Pretesting Questions for In-depth Interviews

Once an interview guide was drafted, two preliminary interviews were conducted to pretest the guide with key informants or members of the target audience.

When pretesting interview guides, ask the following questions:

- Are the questions understandable?
- Do the questions sound awkward?
- Are any questions leading?
- Are any questions closed-ended?
- Do informants understand the questions?
- Do the questions promote discussion?
- Is the guide too long?
- Does the guide address only what is important to the study?
- What works? What does not?
- Are you obtaining the responses you need?

Appendix D: The In-depth Interview Checklist

1. **The interview checklist:**
   - Name and contact information for informant
   - The recorder
   - Pens
   - Paper
   - Incentives (PolyU souvenir)
   - Copy of interview guide
   - Copy of instructions used for the 1st and 2nd round Delphi survey
   - Laptop
2. **Guidelines for interviewers:**

Do	Don't
Give informants sufficient time to speak, or spend more time listening than talking.	Interrupt respondents unless necessary.
Offer reassurance that responses are confidential.	Interview in an environment where there are external interruptions or competing distractions.
Introduce new topics as appropriate.	Put respondents' remarks into your own words.
Adopt a curious, sympathetic attitude.	Express shock or surprise at a response.
Use probing and follow-up questions to solicit responses that are more detailed. – Probe on the last remark made by the respondent. – Probe on an idea expressed earlier in the interview.	Accept the response "I don't know".
Ask for clarification of colloquial or unfamiliar terms.	Give the impression that responses are right or wrong.
Ask for clarification if you do not understand a response.	Embarrass informants by insisting they respond to questions that make them feel uncomfortable.
Use nonverbal cues (e.g. nodding head) and silence as prompts.	Ask "why".
Clearly know what information to find out.	Use words or phrases the respondent will not understand.
Ask the right questions to get the required information.	Spend interview time on irrelevant or unrelated topics, or introduce your own perspective into the interview.
Be familiar with the interview guide to ensure to move back and forth through it as needed.	Jump from one subject to another.
Use encouraging sounds.	Ask leading questions.

Note: Adapted from Longsfield (2004) and The Wallace Foundation (2012).

Appendix E: Questionnaire for Feedback Survey in the Pilot Study

The operation system in your computer: *

- Windows XP
- Windows 7
- Windows Vista
- Mac OS X
- iOS (iPhone)
- Linux
- Windows 1998/2000
- Other:

Which web browser do you use most frequently *

- Internet Explorer
- Firefox
- Google Chrome
- Safari
- Opera
- Other:

Page 2

After page 1
Continue to next page

Self-rating of expertise in tourism demand forecasting: *

	1	2	3	4	5	6	7	
Very little								Excellent

How long did it take you to complete the survey? *

- < 20 minutes
- 21-30 minutes
- 31-40 minutes
- 41-50 minutes
- 51-60 minutes
- > 60 minutes

I have completed the Delphi survey before receiving the reminder email. *

- Yes
- No

Were you able to open and view the video demo? *

- Yes
- No

Page 3

After page 2

Continue to next page

Which instructions did you use? *

- Instructions in WORD
- Video DEMO
- Both
- Neither

Page 4

After page 3

Go to page 8 (To what extent do you a... following statements?)

Note: "Go to page" selections will override this navigation. Learn more.

Use instructions in WORD only.

To what extent do you agree or disagree with the following statements? *

	5 Strongly agree	4 Agree	3 Neutral	2 Disagree	1 Strongly disagree
The instructions provided are clear, accurate, and easy to understand.					
The length of the instructions is appropriate.					
The graphs in the instructions are easy to understand.					
I have a clear understanding of what I am expected to do in the forecast task after reading the instructions.					
The graphs of historical trends of visitor arrivals in the Appendix 1 were useful when I made my adjustments.					
The IMF projections of GDP and exchange rates in the Appendix 2 were useful when I made my adjustments.					

Page 5

After page 4

Go to page 8 (To what extent do you a... following statements?)

Use video demo only.

To what extent do you agree or disagree with the following statements? *

	5 Strongly agree	4 Agree	3 Neutral	2 Disagree	1 Strongly disagree
The instructions provided are clear, accurate, and easy to understand.					
The length of the instructions is appropriate.					
The graphs in the instructions are easy to understand.					
I have a clear understanding of what I am expected to do in the forecast task after watching the video demo.					
The graphs of historical trends of visitor arrivals in the Appendix 1 were useful when I made my adjustments.					
The IMF projections of GDP and exchange rates in the Appendix 2 were useful when I made my adjustments.					

Page 6

After page 5

Go to page 8 (To what extent do you a... following statements?)

Use both WORD and DEMO version.

Please select whether you agree or disagree with the following statements: *

	5 Strongly agree	4 Agree	3 Neutral	2 Disagree	1 Strongly disagree
The instructions (WORD version) provided are clear, accurate, and easy to understand.					
The instructions (VIDEO DEMO version) provided are clear, accurate, and easy					

	5 Strongly agree	4 Agree	3 Neutral	2 Disagree	1 Strongly disagree
to understand.					
The length of the instructions (WORD version) is appropriate.					
The length of the instructions (VIDEO DEMO version) is appropriate.					
The graphs in the instructions are easy to understand.					
I have a clear understanding of what I am expected to do in the forecast task after reading the instructions.					
I have a clear understanding of what I am expected to do in the forecast task after watching the video demo.					
The graphs of historical trends of visitor arrivals in the Appendix 1 were useful when I made my adjustments.					
The IMF projections of GDP and exchange rates in the Appendix 2 were useful when I made my adjustments.					

Page 7

After page 6

Go to page 8 (To what extent do you a... following statements?)

I did not use any instructions.

Why did you not use the instructions? *

Page 8

After page 7

Continue to next page

To what extent do you agree or disagree with the following statements?

Note: 1:Strongly disagree, 2:Disagree, 3:Neutral, 4: Agree, 5:Strongly agree

The forecasting system is easy to use. *

	1	2	3	4	5	
Strongly disagree						Strongly agree

The graphs in the website are more informative than tables. *

	1	2	3	4	5	
Strongly disagree						Strongly agree

The time to complete the forecast tasks is appropriate. *

	1	2	3	4	5	
Strongly disagree						Strongly agree

The button "All experts' adjustment" was useful when I adjusted my forecasts in the 2nd round. *

	1	2	3	4	5	
Strongly disagree						Strongly agree

Page 9

After page 8

Continue to next page

To what extent do you agree or disagree with the following statements?

Note: 1:Strongly disagree, 2:Disagree, 3:Neutral, 4: Agree, 5:Strongly agree

The summarized group forecasts (i.e. mean, median, min, max, quartiles) were useful when I adjusted my forecasts in the 2nd round. *

	1	2	3	4	5	
Strongly disagree						Strongly agree

The graphs in the summary report for the 1st round were useful when I adjusted my forecasts in the 2nd round. *

	1	2	3	4	5	
Strongly disagree						Strongly agree

The tables in the summary report for the 1st round were useful when I adjusted my forecasts in the 2nd round. *

	1	2	3	4	5	
Strongly disagree						Strongly agree

The historical data points (starting from 2006Q1) are sufficient enough to assist me with the adjustments. *

	1	2	3	4	5	
Strongly						Strongly

disagree agree

Which you will suggest as the starting year of presenting historical data? *

- 15 years earlier
- 10 years earlier
- 5 years earlier
- 3 years earlier
- Other:

Page 10

After page 9
Continue to next page

Was there anything missing in the instructions? (Optional)

Any further comments/suggestions? (Optional)

Appendix F: Outputs of Tests of Normality and Homogeneity of Variance

- The first round (R1):

Tests of Normality

market		group	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
			Statistic	df	Sig.	Statistic	df	Sig.
All	mape1	Industry	.242	7	.200*	.914	7	.426
		Academic	.174	11	.200*	.955	11	.711
	rmspe1	Industry	.236	7	.200*	.933	7	.579
		Academic	.162	11	.200*	.985	11	.988
Australia	mape1	Industry	.291	7	.075	.785	7	.029
		Academic	.232	11	.101	.831	11	.024
	rmspe1	Industry	.307	7	.046	.739	7	.010
		Academic	.271	11	.023	.812	11	.014
China	mape1	Industry	.188	7	.200*	.959	7	.811
		Academic	.146	11	.200*	.939	11	.514
	rmspe1	Industry	.185	7	.200*	.960	7	.818
		Academic	.142	11	.200*	.944	11	.573
Japan	mape1	Industry	.240	7	.200*	.910	7	.394
		Academic	.263	11	.032	.721	11	.001
	rmspe1	Industry	.221	7	.200*	.922	7	.485
		Academic	.261	11	.035	.684	11	.000
Taiwan	mape1	Industry	.440	7	.000	.520	7	.000
		Academic	.305	11	.005	.765	11	.003
	rmspe1	Industry	.432	7	.000	.540	7	.000
		Academic	.221	11	.140	.796	11	.008
UK	mape1	Industry	.300	7	.057	.873	7	.197
		Academic	.395	11	.000	.760	11	.003
	rmspe1	Industry	.308	7	.044	.854	7	.133
		Academic	.383	11	.000	.754	11	.002
USA	mape1	Industry	.311	7	.039	.854	7	.134
		Academic	.281	11	.015	.832	11	.024
	rmspe1	Industry	.327	7	.023	.830	7	.080
		Academic	.259	11	.038	.850	11	.043

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

Test of Homogeneity of Variance

market			Levene Statistic	df1	df2	Sig.
All	mape1	Based on Mean	3.178	1	16	.094
		Based on Median	2.003	1	16	.176
		Based on Median and with adjusted df	2.003	1	14.286	.178
		Based on trimmed mean	3.177	1	16	.094
	rmspe1	Based on Mean	1.332	1	16	.265
		Based on Median	1.207	1	16	.288
		Based on Median and with adjusted df	1.207	1	14.890	.289
		Based on trimmed mean	1.340	1	16	.264
Australia	mape1	Based on Mean	.063	1	16	.806
		Based on Median	.063	1	16	.805
		Based on Median and with adjusted df	.063	1	15.942	.805
		Based on trimmed mean	.066	1	16	.801
	rmspe1	Based on Mean	.003	1	16	.957
		Based on Median	.037	1	16	.849
		Based on Median and with adjusted df	.037	1	15.869	.849
		Based on trimmed mean	.008	1	16	.929
China	mape1	Based on Mean	.414	1	16	.529
		Based on Median	.425	1	16	.524
		Based on Median and with adjusted df	.425	1	15.931	.524
		Based on trimmed mean	.406	1	16	.533
	rmspe1	Based on Mean	.233	1	16	.636
		Based on Median	.227	1	16	.640
		Based on Median and with adjusted df	.227	1	15.905	.640
		Based on trimmed mean	.228	1	16	.639
Japan	mape1	Based on Mean	2.965	1	16	.104
		Based on Median	1.598	1	16	.224
		Based on Median and with adjusted df	1.598	1	10.514	.233
		Based on trimmed mean	2.123	1	16	.164
	rmspe1	Based on Mean	2.230	1	16	.155
		Based on Median	1.269	1	16	.277
		Based on Median and with adjusted df	1.269	1	10.453	.285
		Based on trimmed mean	1.729	1	16	.207
Taiwan	mape1	Based on Mean	.055	1	16	.817
		Based on Median	.264	1	16	.614
		Based on Median and with adjusted df	.264	1	15.060	.615
		Based on trimmed mean	.087	1	16	.772
	rmspe1	Based on Mean	.059	1	16	.811
		Based on Median	.345	1	16	.565
		Based on Median and with adjusted df	.345	1	14.900	.566
		Based on trimmed mean	.124	1	16	.729
UK	mape1	Based on Mean	3.213	1	16	.092
		Based on Median	.584	1	16	.456
		Based on Median and with adjusted df	.584	1	12.189	.459
		Based on trimmed mean	2.520	1	16	.132
	rmspe1	Based on Mean	3.412	1	16	.083
		Based on Median	.640	1	16	.436
		Based on Median and with adjusted df	.640	1	12.114	.439
		Based on trimmed mean	2.648	1	16	.123
USA	mape1	Based on Mean	2.730	1	16	.118
		Based on Median	1.235	1	16	.283
		Based on Median and with adjusted df	1.235	1	13.311	.286
		Based on trimmed mean	1.974	1	16	.179
	rmspe1	Based on Mean	1.925	1	16	.184
		Based on Median	1.137	1	16	.302
		Based on Median and with adjusted df	1.137	1	14.136	.304
		Based on trimmed mean	1.374	1	16	.258

- The second round (R2):

Tests of Normality

market		group	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
			Statistic	df	Sig.	Statistic	df	Sig.
All	mape2	Industry	.279	6	.156	.760	6	.025
		Academic	.206	11	.200*	.827	11	.021
	rmspe2	Industry	.319	6	.057	.809	6	.070
		Academic	.372	11	.000	.583	11	.000
Australia	mape2	Industry	.448	6	.000	.658	6	.002
		Academic	.291	11	.010	.859	11	.056
	rmspe2	Industry	.481	6	.000	.539	6	.000
		Academic	.324	11	.002	.845	11	.037
China	mape2	Industry	.277	6	.167	.822	6	.092
		Academic	.309	11	.004	.736	11	.001
	rmspe2	Industry	.260	6	.200*	.794	6	.051
		Academic	.289	11	.011	.712	11	.001
Japan	mape2	Industry	.407	6	.002	.672	6	.003
		Academic	.408	11	.000	.525	11	.000
	rmspe2	Industry	.406	6	.002	.676	6	.003
		Academic	.394	11	.000	.523	11	.000
Taiwan	mape2	Industry	.339	6	.030	.826	6	.100
		Academic	.402	11	.000	.673	11	.000
	rmspe2	Industry	.358	6	.016	.819	6	.087
		Academic	.401	11	.000	.719	11	.001
UK	mape2	Industry	.492	6	.000	.496	6	.000
		Academic	.422	11	.000	.592	11	.000
	rmspe2	Industry	.492	6	.000	.496	6	.000
		Academic	.385	11	.000	.547	11	.000
USA	mape2	Industry	.492	6	.000	.496	6	.000
		Academic	.329	11	.002	.853	11	.047
	rmspe2	Industry	.492	6	.000	.496	6	.000
		Academic	.351	11	.000	.809	11	.012

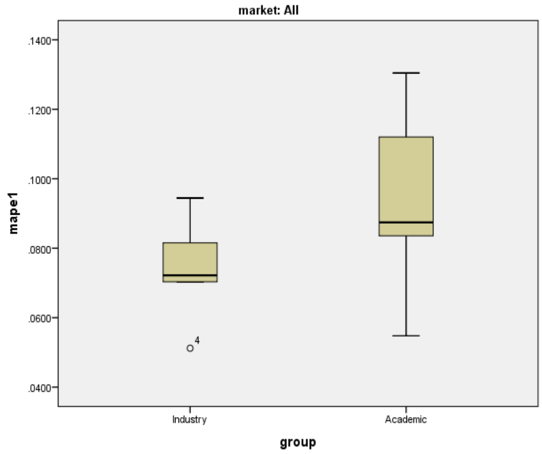
*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

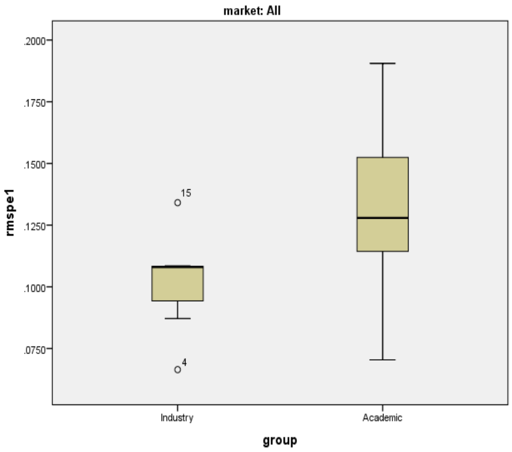
Test of Homogeneity of Variance

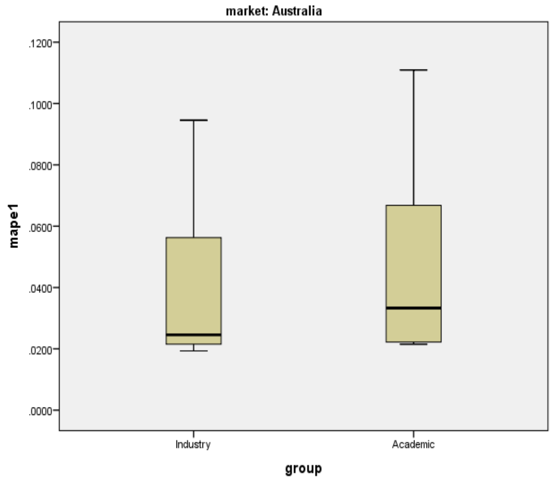
market			Levene Statistic	df1	df2	Sig.
All	mape2	Based on Mean	.013	1	15	.912
		Based on Median	.000	1	15	.996
		Based on Median and with adjusted df	.000	1	14.668	.996
		Based on trimmed mean	.013	1	15	.912
	rmspe2	Based on Mean	.327	1	15	.576
		Based on Median	.158	1	15	.696
		Based on Median and with adjusted df	.158	1	11.616	.698
		Based on trimmed mean	.104	1	15	.751
Australia	mape2	Based on Mean	1.758	1	15	.205
		Based on Median	.250	1	15	.624
		Based on Median and with adjusted df	.250	1	7.919	.630
		Based on trimmed mean	1.154	1	15	.300
	rmspe2	Based on Mean	3.317	1	15	.089
		Based on Median	.342	1	15	.567
		Based on Median and with adjusted df	.342	1	6.716	.578
		Based on trimmed mean	2.044	1	15	.173
China	mape2	Based on Mean	1.081	1	15	.315
		Based on Median	1.100	1	15	.311
		Based on Median and with adjusted df	1.100	1	14.952	.311
		Based on trimmed mean	1.156	1	15	.299
	rmspe2	Based on Mean	1.442	1	15	.248
		Based on Median	.943	1	15	.347
		Based on Median and with adjusted df	.943	1	13.741	.348
		Based on trimmed mean	1.262	1	15	.279
Japan	mape2	Based on Mean	1.873	1	15	.191
		Based on Median	.629	1	15	.440
		Based on Median and with adjusted df	.629	1	10.238	.446
		Based on trimmed mean	1.079	1	15	.315
	rmspe2	Based on Mean	1.674	1	15	.215
		Based on Median	.593	1	15	.453
		Based on Median and with adjusted df	.593	1	10.273	.459
		Based on trimmed mean	.954	1	15	.344
Taiwan	mape2	Based on Mean	4.841	1	15	.044
		Based on Median	.928	1	15	.351
		Based on Median and with adjusted df	.928	1	10.601	.357
		Based on trimmed mean	3.855	1	15	.068
	rmspe2	Based on Mean	4.791	1	15	.045
		Based on Median	.926	1	15	.351
		Based on Median and with adjusted df	.926	1	10.793	.357
		Based on trimmed mean	3.781	1	15	.071
UK	mape2	Based on Mean	.041	1	15	.843
		Based on Median	.003	1	15	.955
		Based on Median and with adjusted df	.003	1	14.286	.955
		Based on trimmed mean	.006	1	15	.941
	rmspe2	Based on Mean	.057	1	15	.814
		Based on Median	.047	1	15	.831
		Based on Median and with adjusted df	.047	1	14.977	.831
		Based on trimmed mean	.053	1	15	.822
USA	mape2	Based on Mean	10.566	1	15	.005
		Based on Median	2.511	1	15	.134
		Based on Median and with adjusted df	2.511	1	11.925	.139
		Based on trimmed mean	9.935	1	15	.007
	rmspe2	Based on Mean	24.992	1	15	.000
		Based on Median	3.447	1	15	.083
		Based on Median and with adjusted df	3.447	1	10.172	.093
		Based on trimmed mean	21.488	1	15	.000

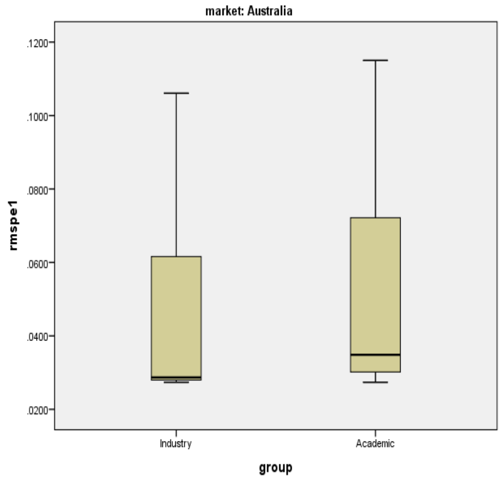
Appendix G: Boxplots of MAPEs and RMSPEs for Two Expert Groups

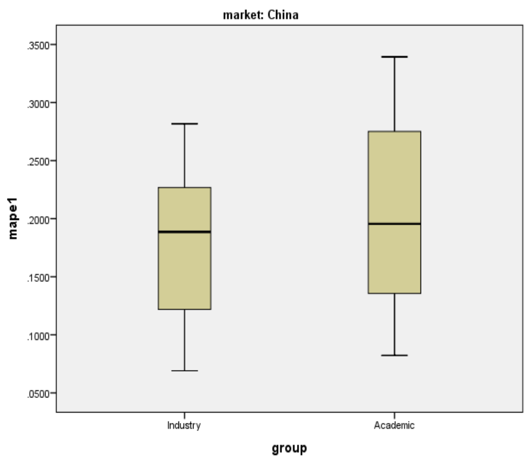
- In the first round (R1)

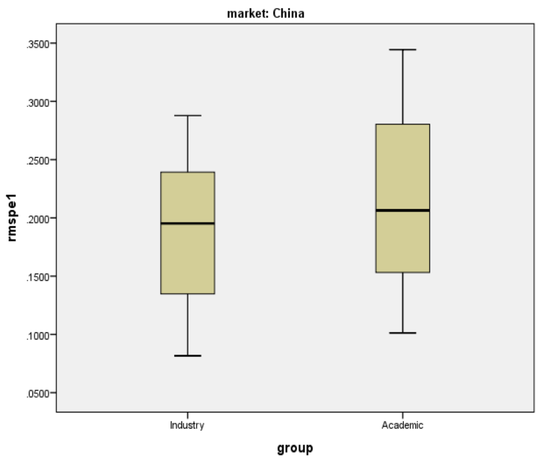


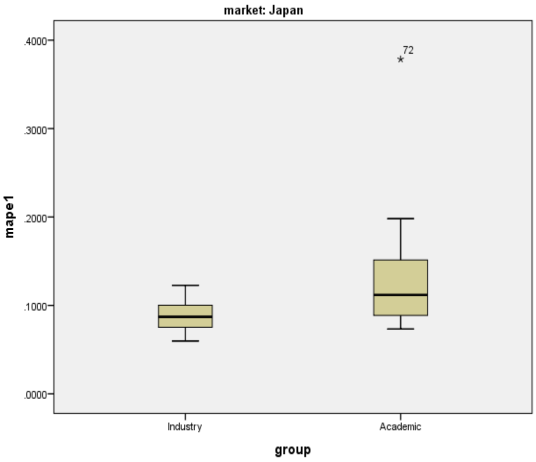


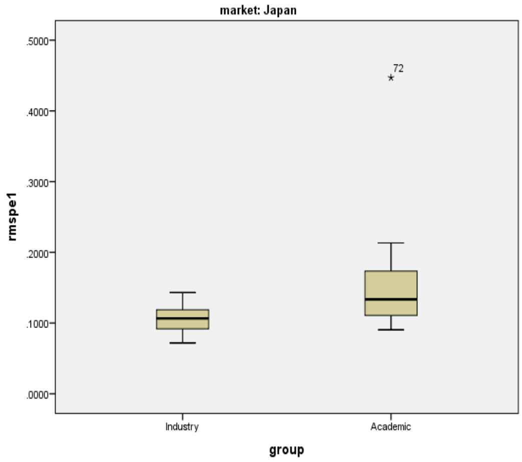


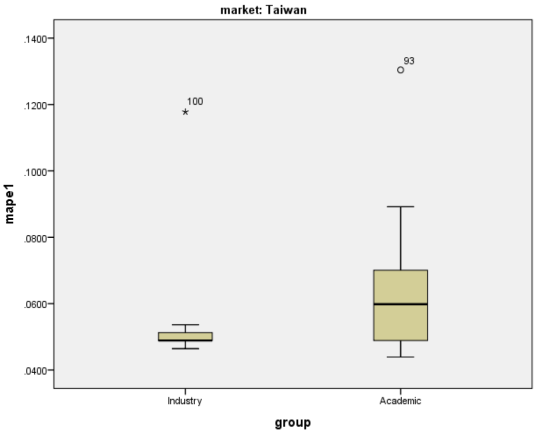


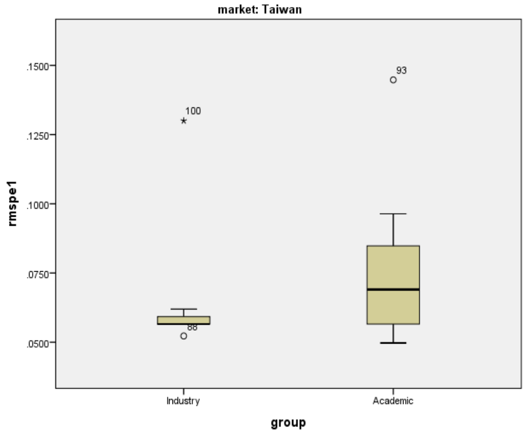


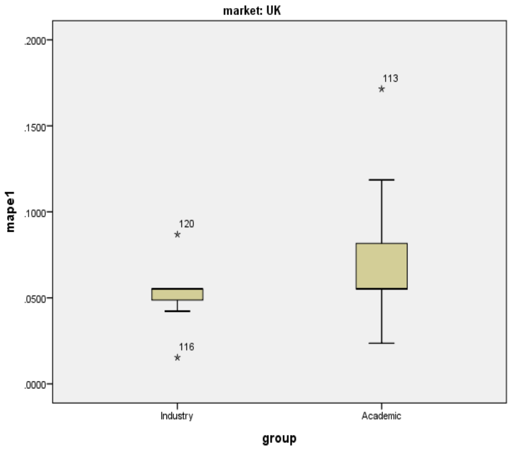


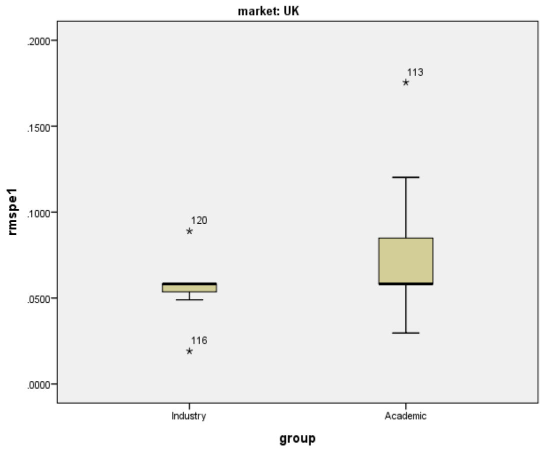

market: Japan
.4000
.3000
.2000
.1000
.0000
mape1
72
Industry
Academic
group

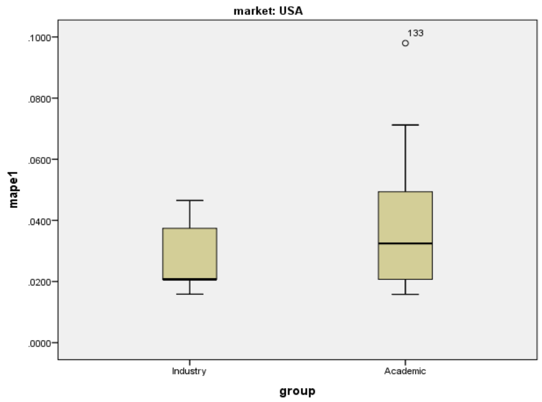

market: Japan
.5000
.4000
.3000
.2000
.1000
.0000
rmspe1
72
Industry
Academic
group

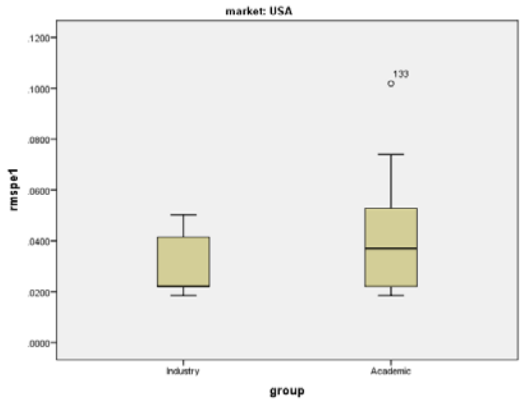

market: Taiwan
.1400
.1200
.1000
.0800
.0600
.0400
mape1
100
93
Industry
Academic
group


market: Taiwan
.1500
.1250
.1000
.0750
.0500
rmspe1
100
93
88
Industry
Academic
group

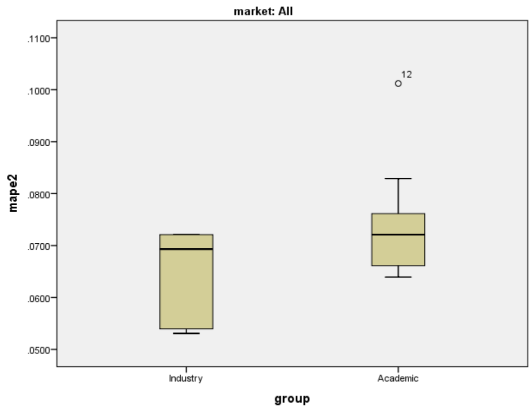

market: UK
.2000
.1500
.1000
.0500
.0000
mape1
113
120
116
Industry
Academic
group

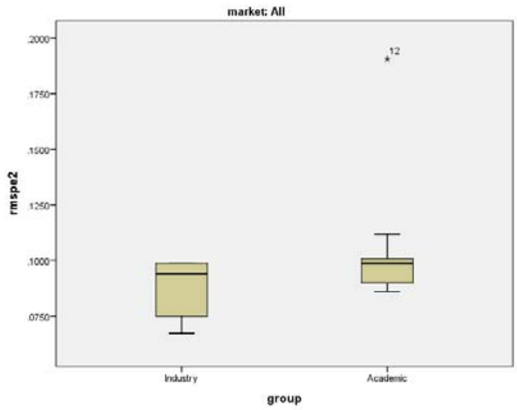

market: UK
.2000
.1500
.1000
.0500
.0000
rmspe1
113
120
116
Industry
Academic
group

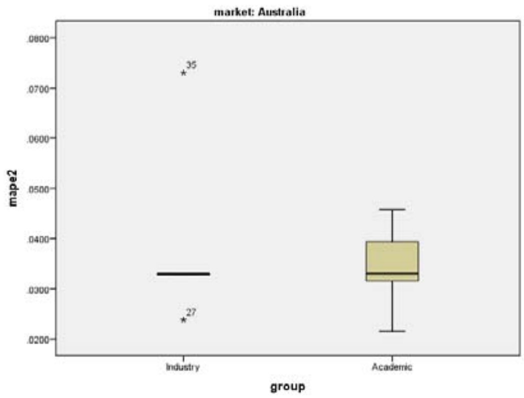


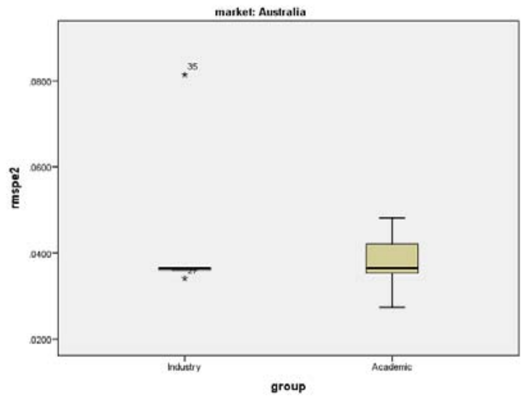


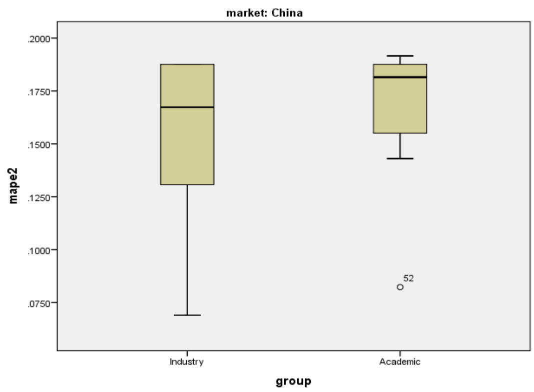
- In the second round (R2)

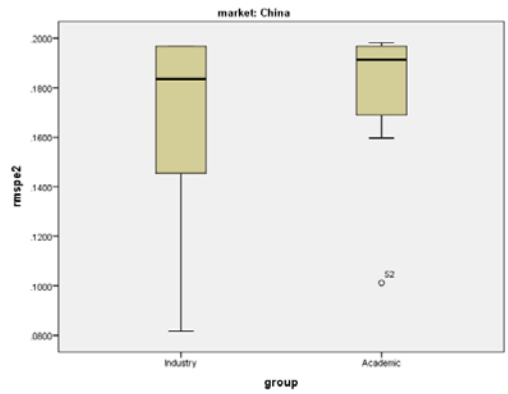


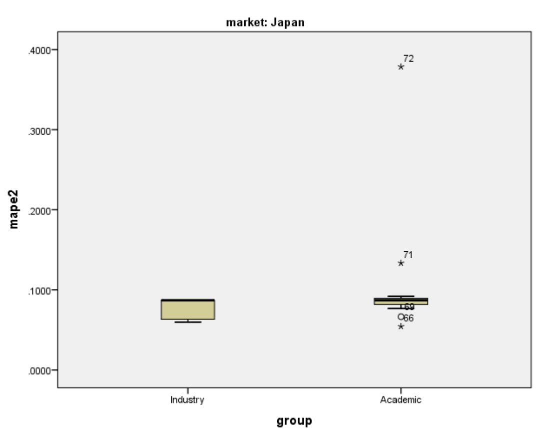


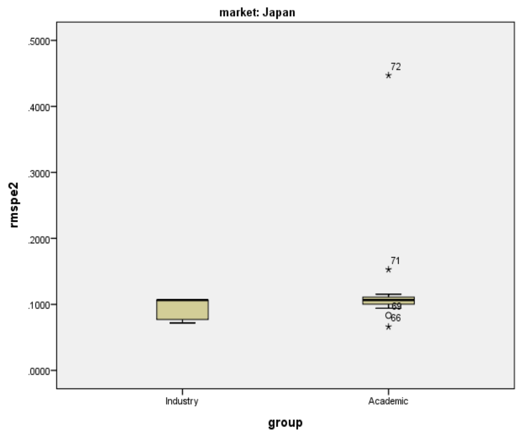


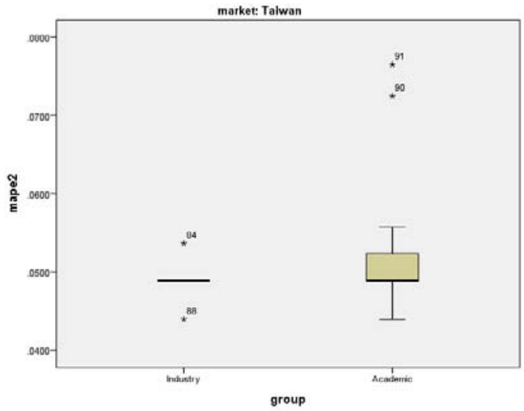


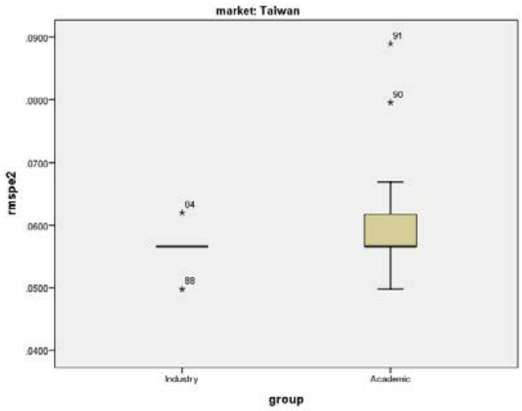


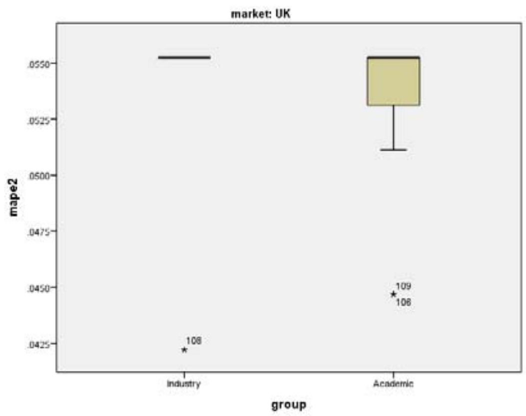


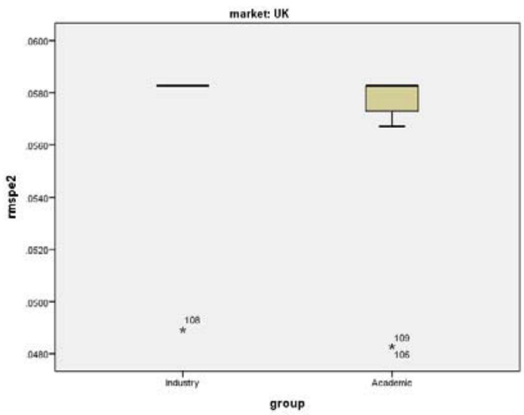

market: Japan
mape2
Industry
Academic
group

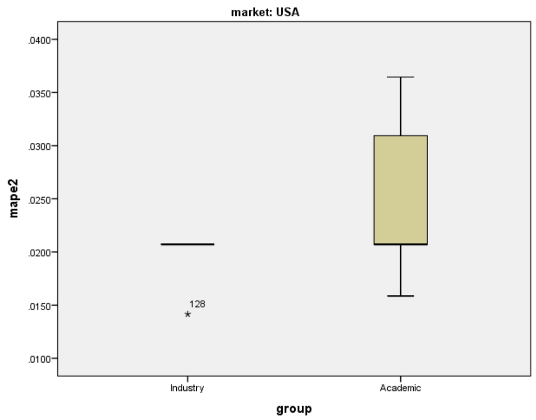

market: Japan
rmspe2
Industry
Academic
group

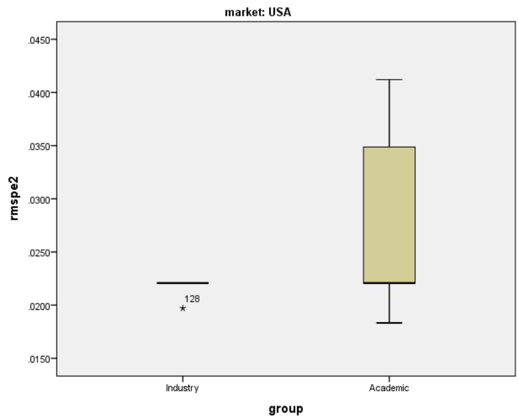

market: Taiwan
mape2
Industry
Academic
group

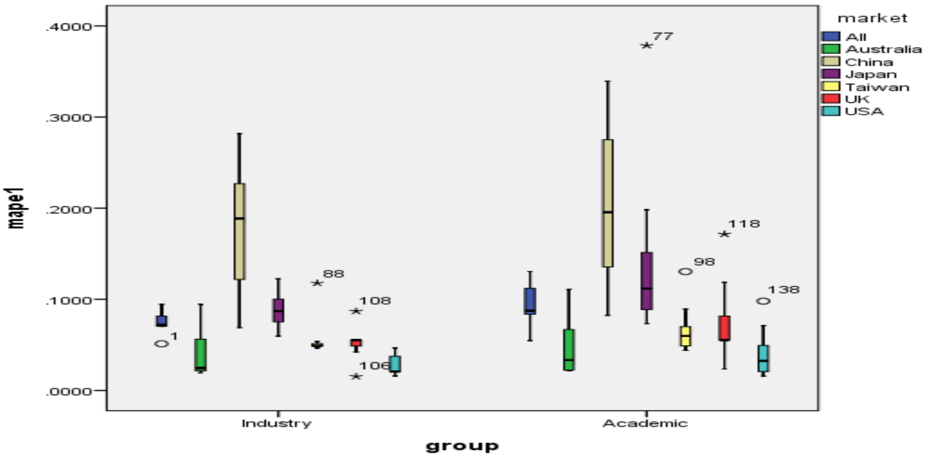

market: Taiwan
rmspe2
Industry
Academic
group

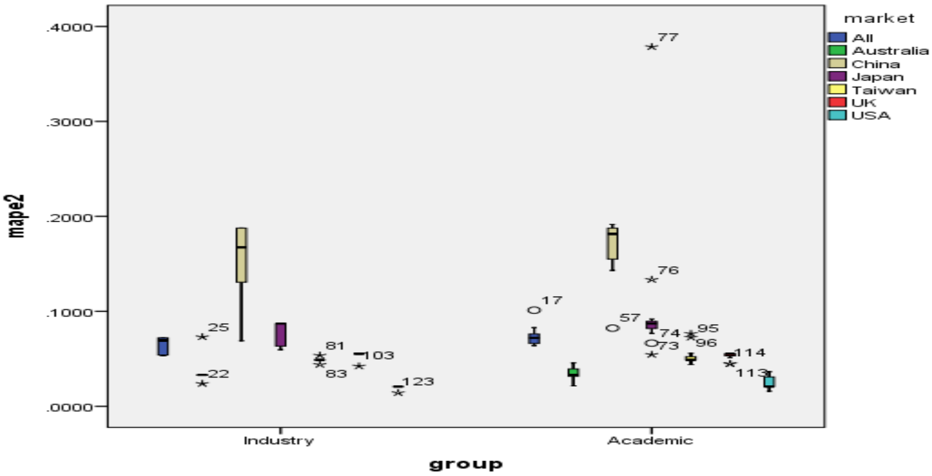

market: UK
mape2
Industry
Academic
group

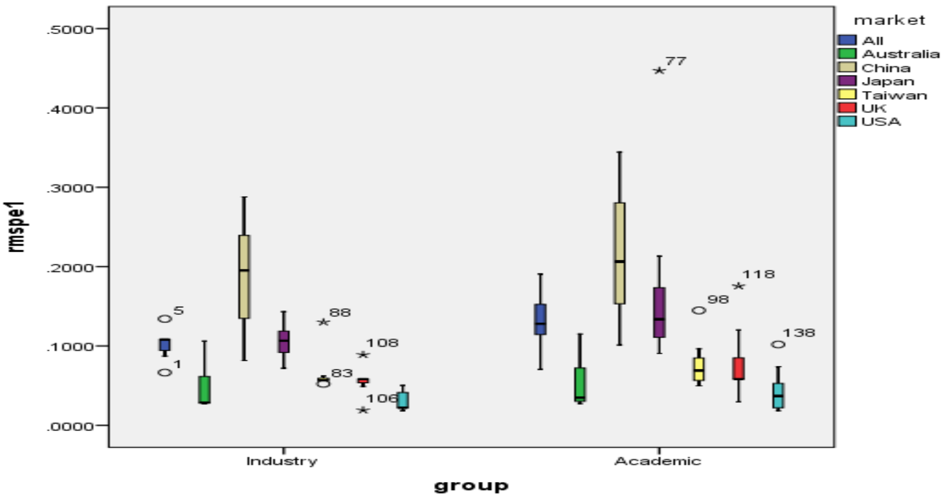

market: UK
rmspe2
Industry
Academic
group

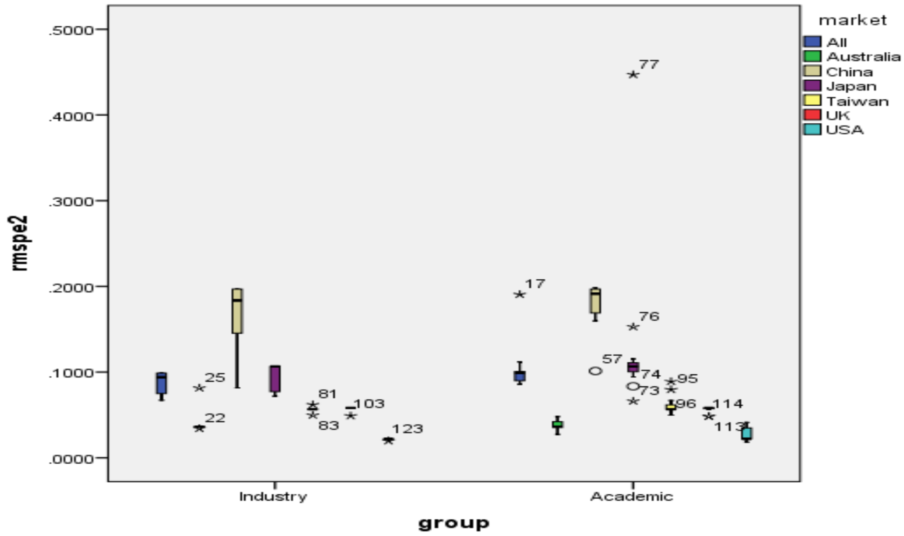

market: USA
mape2
Industry
Academic
group


market: USA
rmspe2
Industry
Academic
group


market
All
Australia
China
Japan
Taiwan
UK
USA
mape1
.4000
.3000
.2000
.1000
.0000
Industry
Academic
group
77
118
98
138
88
108
1
106


market
All
Australia
China
Japan
Taiwan
UK
USA
mape2
.4000
.3000
.2000
.1000
.0000
Industry
Academic
group
77
76
17
57
74
95
73
96
114
113
25
22
81
83
103
123


market
All
Australia
China
Japan
Taiwan
UK
USA
rmspe1
.5000
.4000
.3000
.2000
.1000
.0000
Industry
Academic
group
77
118
98
138
5
1
88
108
83
106


market
All
Australia
China
Japan
Taiwan
UK
USA
rmspe2
.5000
.4000
.3000
.2000
.1000
.0000
77
17
76
57
74
73
95
96
114
113
25
22
81
83
103
123
Industry
Academic
group

References

Ahlburg, D. A. (1984). Forecast evaluation and improvement using Theils decomposition. *Journal of Forecasting*, *3* (3), 345-351.

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.

Allcock, J. (1989). Seasonality. In S. F. Witt, & L. Moutinho, *Tourism marketing and management handbook* (pp. 387-392). London: Prentice Hall.

Anderson, D. R., Burnham, K. P., & White, G. C. (1998).Comparison of Akaike information criterionand consistent Akaike information criterionfor model selection and statistical inferencefrom capture and recapture studies. *Journal of Applied Statistics*, *25*(2), 263-282.

Angus-Leppan, P., & Fatseas, V. (1986). The forecasting accuracy of trainee accountants using judgmental and statistical techniques. *Accounting and Business Research*, *16*, 179-188.

Archer, B. (2000). Demand forecasting and estimation. In C. Tisdell, *The economics of tourism* (Vol. 1, pp. 61-68). Cheltenham: Edward Elgar Pub.

Archer, B. (1976). *Demand forecasting in tourism.* Bangor: University of Wales Press.

Archer, B. (1980). Forecasting demand: Quantitative and intuitive techniques. *International Journal of Tourism Management*, *1* (1), 5-12.

Arkes, H. R. (2001). Overconfidence in judgmental forecasting. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 495-515). Dordrecht: Kluwer Academic.

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

Armstrong, J. (2001b). Evaluating forecasting methods. In J. Armstrong, *Principles of forecasting: A handbook for researchers and practitioners* (pp. 443-472). Dordrecht: Kluwer Academic.

Armstrong, J. (2001c). Judgmental bootstrapping: Inferring experts rules for forecasting. In J. S. Armstrong, *Principles of forecasting: A handbook for researchers and practitioners.* Dordrecht: Kluwer Academic.

Armstrong, J. (2001d). *Principles of forecasting: A handbook for researchers and practitioners.* Dordrecht: Kluwer Academic.

Armstrong, J. (2001e). Selecting forecasting methods. In J. Armstrong, *Principles of forecasting: A handbook for researchers and practitioners.* Dordrecht, MA: Kluwer Academic.

Armstrong, J. (1985). *Long-range forecasting: From crystal ball to computer.* New York: Wiley.

Armstrong, J. (2006). Findings from evidence-based forecasting: Methods for reducing forecast error. *International Journal of Forecasting*, *22* (3), 583-598.

Armstrong, J., & Collopy, F. (1993). Causal forces: Structuring knowledge for time series extrapolation. *Journal of Forecasting*, *12*, 103-115.

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.

Armstrong, J., & Lusk, E. J. (1983). Commentary on the Makridakis time series competition (M-Competition). *Journal of Forecasting*, *2*, 259-311.

Armstrong, J., Denniston, W. B., & Gordon, M. M. (1975). The use of the decomposition principle in making judgments. *Organizational Behavior and Human Performance*, *14*, 257-263.

Asch, S. E. (2003). Effects of group pressure upon the modification and distortion of judgments. In L. W. Porter, H. L. Angle, & R. W. Allen, *Organizational influence processes* (pp. 295-303). New York: M. E. Sharpe.

Ashton, A. H., & Ashton, R. H. (1985). Aggregating subjective forecasts: Some empirical results. *Management Science*, *31*, 1499-1508.

Austin, D., Leeb, Y., & Getzb, D. (2008). A Delphi study of trends in special and inclusive recreation. *Leisure/Loisir*, *32* (1), 163-182.

BarOn, R. (1975). *Seasonality in tourism: A guide to the analysis of seasonality and trends for policy making.* London, UK: Economist Intelligence Unit.

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Operational Research Quarterly*, *20* (4), 451-468.

Beach, L., & Christensen-Szalanski, J. (1987). Assessing human judgment: Has it been done, can it be done, should it be done? In G. Wright, & P. Ayton, *Judgmental forecasting* (pp. 49-62). New York: Wiley.

Benbasat, I., & Dexter, A. S. (1985). An experimental evaluation of graphical and color-enhanced information presentation. *Management Science*, *31* (11), 1348-1364.

Benson, P. G., & Önkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*, *8* (4), 559-573.

Bessler, D. A., & Chamberlain, P. (1987). On Bayesian composite forecasting. *Omega*, *15*, 43-48.

Best, R. J. (1974). An experiment in Delphi estimation in marketing decision making. *Journal of Marketing Research*, *11* (4), 448-452.

Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*, *36* (8), 887-899.

Bonham, C., Gangnes, B., & Zhou, T. (2009). Modeling tourism: A fully identified VECM approach. *International Journal of Forecasting*, *25* (3), 531-549.

Box, G. E., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control.* San Francisco: Holden-Day.

Boyce, C., & Neale, P. (2006). *Conducting in-depth interviews: A guide for designing and conducting in-depth interviews for evaluation input.* Retrieved July 25, 2012, from http://www2.pathfinder.org/site/DocServer/m_e_tool_series_indepth_interviews.pdf

Bulter, R. (2001). Seasonality in tourism: Issues and implications. In T. Baum, & S. Lundtorp, *Seasonality in tourism.* Oxford: Elsevier.

Bunn, D. (1975). A Bayesian approach to the linear combination of forecasts. *Operational Research Quarterly*, *26* (2), 325-329.

Bunn, D. (1981). Two methodologies for the linear combination of forecasts. *Journal of the Operational Researc Society*, *32*, 213-222.

Bunn, D. (1988). Combining forecasts. *European Journal of Operational Research*, *33* (3), 223-229.
Bunn, D., & Wright, G. (1991). Interaction of judgemental and statistical forecasting methods: Issues and analysis. *Management Science*, *37* (5), 501-518.
Burger, C. J., Dohnal, M., Kathrada, M., & Law, R. (2001). A practitioners guide to time-series methods for tourism demand forecasting- A case study of Durban, South Africa. *Tourism Management*, *22* (4), 403-409.
Butler, R. (1994). Seasonality in tourism: Issues and problems. In A. V. Seaton, C. L. Jenkins, R. C. Wood, P. U. Dieke, M. M. Bennett, L. R. Maclellan, et al., *Tourism: The state of the art* (pp. 332-339). Chichester: Wiley.
Butler, W., Kavesh, R. A., & Platt, R. B. (1974). *Methods and techniques of business forecasting.* Englewood Cliffs, N.J.: Prentice-Hall.
Caiden, N., & Wildavsky, A. (1974). *Planning and budgeting in poor countries.* New York: Wiley.
Calantone, R. J., Benedetto, C. A., & Bojanic, D. (1987). A comprehensive review of the tourism forecasting literature. *Journal of Travel Research*, *26* (2), 28-39.
Caner, M., & Kilian, L. (2001). Size distortions of tests of the null hypothesis of stationarity: Evidence and implications for the PPP debate. *Journal of International Money and Finance*, *20*, 639-657.
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.
CB. (2011). *Consumer confidence survey technical note – February 2011.* Retrieved November 20, 2011, from The Conference Board (CB): http://www.conference-board.org/data/consumerconfidence.cfm.
Census and Statistics Department. (2011). *Hong Kong annual digest of statistics.* Retrieved August 1, 2012, from Census and Statistics Department, Hong Kong, SAR: http://www.statistics.gov.hk/pub/B10100032011AN11B0100.pdf
Chakravarti, D., Mitchell, A., & Staelin, R. (1979). Judgment based marketing decision models: An experimental investigation of the decision calculus approach. *Management Science*, *25* (3), 251-263.
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.
Chung, J. Y. (2009). Seasonality in tourism: A review. *e-Review of Tourism Research (eRTR)*, *7* (5), 82-96.
Clemen, R. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5* (4), 559-583.
Collopy, F., & Armstrong, J. (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.
Creswell, J. W., & Clark, V. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.
Croce, V., & Wöber, K. W. (2011). Judgemental forecasting support systems in tourism. *Tourism Economics*, *17* (4), 709-724.

Crouch, G. I. (1994). The study of international tourism demand: A survey of practice. *Journal of Travel Research*, *32*, 41-54.
Cunliffe, S. (2002). *Forecasting risks in the tourism industry using the Delphi technique.* Retrieved from Tourism: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.5350&rep=rep1&type=pdf
D'Amore, L. (1976). The significance of tourism to Canada. *Business Quarterly, 41* (3), 227-235.
Dalkey, N. C. (1969). *Delphi method: An experimental study of group opinion.* US: The Rand Corporation.
Dalkey, N. C., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science*, *9* (3), 458-467.
Dalkey, N. C., Brown, B., & Cochran, S. (1969). *The Delphi method, III: Use of self ratings to improve group estimates.* Retrieved August 21, 2012, from The Rand corporation: http://www.rand.org/pubs/research_memoranda/2006/RM6115.pdf
Dalrymple, D. J. (1975). Sales forecasting methods and accuracy. *Business Horizons*, *18* (6), 69-73.
Dalrymple, D. J. (1987). Sales forecasting practices: Results from a United States survey. *International Journal of Forecasting*, *3* (3-4), 379-391.
Day, J., & Bobeva, M. (2005). A generic toolkit for the successful management of Delphi studies. *Electronic Journal of Business Research Methods*, *3* (2), 103-116.
de Menezes, L. M., Bunn, D. W., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, *120* (1), 190-204.
De Vita, G. & Abbott, A. (2002). Are saving and investment cointegrated? An ARDL bounds testing approach. *Economics Letters*, *77*, 293-299.
Delbecq, A., Van de Ven, A., & Gustafson, D. (1975). *Group techniques for program planning: A guide to nominal group and Delphi processes.* Glenview IL: Scott,Foresman.
Denzin, N., & Lincoln, Y. (2000). *Handbook of qualitative research.* London: Sage.
Department of Tourism. (2011). *Tourist arrivals in Thailand.* Retrieved April 8, 2011, from Department of Tourism (Thailand): http://tourism.go.th/2010/en/statistic/tourism.php
Diamantopoulos, A., & Mathews, B. (1989). Factors affecting the nature and effectiveness of subjective revision in sales forecasting: An empirical study. *Managerial & Decision Economics*, *10* (1), 51-59.
Diebold, F. X., & Lopez, J. A. (1996). Forecast evaluation and combination. *SSRN eLibrary*.
Donohoe, H. (2011a). A Delphi toolkit for ecotourism research. *Journal of Ecotourism*, *10* (1), 1-20.
Donohoe, H. (2011b). Defining culturally sensitive ecotourism: A Delphi consensus. *Current Issues in Tourism*, *14* (1), 27-45.
Donohoe, H., & Needham, R. D. (2009). Moving best practice forward: Delphi characteristics, advantages, potential problems, and solutions. *International Journal of Tourism Research*, *11* (5), 415-437.
Dwyer, L., Forsyth, P., & Dwyer, W. (2010). *Tourism economics and policy.* Bristol: Channel View Publications.
Dyck, H. J., & Emery, G. J. (1970). *Social futures: Alberta, 1970-2005.* Edmonton: Human Resources Research Council of Alberta.

Edgell, D. L., Seely, R. L., & Iglarsh, H. J. (1980). Forecasts of international tourism to the USA. *International Journal of Tourism Management*, *1* (2), 109-113.

Edmundson, B., Lawrence, M., & O'Connor, M. (1988). The use of non-time series information in sales forecasting: A case study. *Journal of Forecasting*, *7* (3), 201-211.

Edmundson, R. (1990). Decomposition: A strategy for judgemental forecasting. *Journal of Forecasting*, *9* (4), 305-314.

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

Elgers, P. T., Lo, M. H., & Murray, D. (1995). Note on adjustments to analysts earnings forecasts based upon systematic cross-sectional components of prior-period errors. *Management Science*, *41* (8), 1392-1396.

English, M. J., & Kernan, G. L. (1976). The prediction of air travel and aircraft technology to the year 2000 using the Delphi method. *Transportation Research*, *10*(1), 1-8.

Eroglu, C. (2006). An investigation of accuracy, learning and biases in judgmental adjustments of statistical forecasts. *PhD Dissertation*. Ohio, United States: The Ohio State University.

Eroglu, C., & Knemeyer, A. M. (2010). Exploring the potential effects of forecaster motivational orientation and gender on judgmental adjustments of statistical forecasts. *Journal of Business Logistics*, *31* (1), 179-196.

Evans, G. (1995). Planning for the British millennium festival: Establishing the visitor baseline and a framework for forecasting. *Festival Management and Event Tourism*, *3* (4), 183-196.

FAA. (2004). *FAA Aerospace forecast fiscal years 2004-2015.* Washington, DC: Federal Aviation Administration.

FAA. (2010). *FAA Aerospace forecast fiscal years 2010-2030.* Washington, DC: Federal Aviation Administration.

FAA. (2011). *FAA Aerospace forecast fiscal years 2011-2031.* U.S. Department of Transportation. Washington, DC: Federal Aviation Administration.

Faulkner, B., & Valerio, P. (1995). An integrative approach to tourism demand forecasting. *Tourism Management*, *16* (1), 29-37.

Fazlollahi, B., Parikh, M. A., & Verma, S. (1995). *Evaluation of decisional guidance in decision support systems: An empirical study.* Retrieved November 2, 2012, from Department of Decision Sciences, Georgia State University: http://verma.sfsu.edu/profile/idsi-95.pdf

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Fildes, R. (1985). Quantitative forecasting-The state of the art: Econometric models. *The Journal of the Operational Research Society*, *36* (7), 549-580.

Fildes, R. (1991). Efficient use of information in the formation of subjective industry forecasts. *Journal of Forecasting*, *10* (6), 597-617.

Fildes, R., & Beard, C. (1992). Forecasting systems for production and inventory control. *International Journal of Operations and Production Management*, *12*, 4-27.

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. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems*, *42* (1), 351-361.

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.

Fischer, I., & Harvey, N. (1999). Combining forecasts: What information do judges need to outperform the simple average? *International Journal of Forecasting*, *15* (3), 227-246.

Flores, B. E., & White, E. M. (1988). A framework for the combination of forecasts. *Journal of the Academy of Marketing Science*, *16*, 95-102.

Flores, B. E., & White, E. M. (1989). Subjective versus objective combining of forecasts: An experiment. *Journal of Forecasting*, *8* (3), 331-341.

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.

Fosu, O., & Magnus, F. (2006). Bounds testing approach to cointegration: An examination of foreign direct investment trade and growth relationships. *American Journal of Applied Science*, *3* (11), 2079-2085.

Fourie, J., & Santana-Gallego, M. (2010). *The impact of mega-events on tourist arrivals.* Retrieved June 20, 2010, from http://www.econrsa.org/papers/w_papers/wp171.pdf

Frechtling, D. C. (2001). *Forecasting tourism demand: Methods and strategies.* Oxford: Butterworth-Heinemann.

Fritz, R. G., Brandon, C., & Xander, J. (1984). Combining time-series and econometric forecast of tourism activity. *Annals of Tourism Research*, *11* (2), 219-229.

Garrod, B., & Fyall, A. (2000). Managing heritage tourism. *Annals of Tourism Research, 27* (3), 682-708.

Garrod, B., & Fyall, A. (2005). Revisiting Delphi: The Delphi technique in tourism research. In B. Ritchie, & C. Palmer (Eds.), *Tourism research methods: Integrating theory with practice* (pp. 85-98). Cambridge: CABI.

Getz, D. (2008). Event tourism: Definition, evolution, and research. *Tourism Management*, *29* (3), 403-428.

Ghalia, M. B., & Wang, P. P. (2000). Intelligent system to support judgmental business forecasting: The case of estimating hotel room demand. *IEEE Transactions on Fuzzy Systems*, *8* (4), 380-397.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuritics and biases: The psychology of intuitive judgment.* New York: Cambridge.

Goeldner, C. R., & Ritchie, J. R. (2005). *Tourism: Principles, practices, philosophies.* Hoboken: Wiley.

Goodwin, P. (1996). Statistical correction of judgmental point forecasts and decisions. *Omega-International Journal of Management Science*, *24* (5), 551-559.

Goodwin, P. (2000a). Correct or combine? Mechanically integrating judgmental forecasts with statistical methods. *International Journal of Forecasting*, *16* (2), 261-275.

Goodwin, P. (2000b). Improving the voluntary integration of statistical forecasts and judgment. *International Journal of Forecasting*, *16* (1), 85-99.

Goodwin, P. (2002). Integrating management judgment and statistical methods to improve short-term forecasts. *Omega*, *30* (2), 127-135.

Goodwin, P. (2005). How to integrate management judgment with statistical forecasts. *Foresight: The International Journal of Applied Forecasting*, *1*, 8-12.
Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making*, *12* (1), 37-53.
Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, *15* (4), 405-408.
Goodwin, P., & Wright, G. (1994). Heuristics, biases and improvement strategies in judgmental time-series forecasting. *Omega-International Journal of Management Science*, *22* (6), 553-568.
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.
Goodwin, P., & Wright, G. (2010). The limits of forecasting methods in anticipating rare events. *Technological Forecasting and Social Change*, *77* (3), 355-368.
Gordon, T. (1994). The Delphi method. *Futures Research Methodology*. Retrieved August 10, 2011, from http://www.gerenciamento.ufba.br/Downloads/delphi%20(1).pdf
Granger, C. W., & Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, *3* (2), 197-204.
Green, H., & Hunter, C. (1992). The environmental impact assessment of tourism development. In P. Johnson, & B. Thomas, *Perspectives on tourism policy* (pp. 29-47). London: Mansell.
Green, H., Hunter, C., & Moore, B. (1990a). Application of the Delphi technique in tourism. *Annals of Tourism Research*, *17* (2), 270-279.
Green, H., Hunter, C., & Moore, B. (1990b). Assessing the environmental impact of tourism development: Use of the Delphi technique. *Tourism Management*, *11* (2), 111-120.
Guerard, J. B., & Beidleman, C. R. (1987). Composite earnings forecasting efficiency. *Interfaces*, *17* (5), 103-113.
Guion, L. A., Diehl, D. C., & McDonald, D. (2011). *Conducting an in-depth interview*. Retrieved July 25, 2012, from http://edis.ifas.ufl.edu/pdffiles/FY/FY39300.pdf
Gujarati, D. N., & Porter, D. C. (2002). *Basic econometrics* (4th ed.). McGraw-Hill Irwin.
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/
Hallman, J., & Kamstra, M. (1989). Combining algorithms based on robust estimation techniques and cointegration restrictions. *Journal of Forecasting*, *8* (3), 189-198.
Hardy, M. A. (1993). *Regression with dummy variables (Sage University Paper series on Quantitative Applications in the Social Sciences, series No. 07-093).* Newbury Park, CA: Sage.
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.
Harris, R., & Sollis, R. (2003). *Applied time series modelling and forecasting* (1 ed.). New York: John Wiley & Sons Ltd.

Harvey, N. (2001). Improving judgment in forecasting. In J. S. Armstrong, *Principles of forecasting: A handbook for researchers and practitioners* (pp. 59-80). Norwell, MA: Kluwer Academic Publishers.
Harvey, N. (2007). Use of heuristics: Insights from forecasting research. *Thinking & Reasoning*, *13* (1), 5-24.
Harvey, N., & Bolger, F. (1996). Graphs versus tables: Effects of data presentation format on judgemental forecasting. *International Journal of Forecasting*, *12* (1), 119-137.
Harvey, N., & Harries, C. (2004). Effects of judges forecasting on their later combination of forecasts for the same outcomes. *International Journal of Forecasting*, *20* (3), 391-409.
Hawkins, D., Shafer, E., & Rovelstad, J. (1980). *Summary and recommendations: International symposium on tourism and the next decade.* Washington, D.C.: George Washington University.
Henderson, K., & Bialeschki, M. (1984). Organized camping and the future: Research on major trends. *Camping Magazine*, *56* (3), 20-26.
Hendry, D. (1980). 'Econometrics-Alchemy or science?'. *Economica*, *47* (188), 387-406.
Hendry, D. (1995). *Dynamic econometrics: An advanced text in econometrics.* Oxford: Oxford University Press.
Hibon, M., & Evgeniou, T. (2005). To combine or not to combine: Selecting among forecasts and their combinations. *International Journal of Forecasting*, *21* (1), 15-24.
Hill, K. Q., & Fowles, J. (1975). The methodological worth of the Delphi technique. *Technological Forecasting and Social Change*, *7*, 179-192.
HKTB. (2011). *Visitor arrival statistics (1985-2011).* Retrieved April 3, 2011, from Hong Kong Tourism Board: http://partnernet.hktb.com/
HKTB. (2012a). *Visitor arrival statistics.* Retrieved September 20, 2012, from Hong Kong Tourism Board: http://partnernet.hktb.com/
HKTB. (2012b). *Tourism expenditure associated to inbound tourism.* Retrieved September 20, 2012, 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: http://www.tourismforecasting.net/hktdfs/
Hogarth, R. (1985). *Judgement and choice: The psychology of decision* (2nd ed.). New York: Wiley.
Horne, J., & Manzenreiter, W. (2004). Accounting for mega-events: forecast and actual impacts of the 2002 Football World Cup Finals on the host countries Japan/Korea. *International Review for the Sociology of Sport*, *39* (2), 187-203.
Hsu, C., & Sandford, B. (2007). The Delphi technique: Making sense of consensus. *Practical Assessment, Research & Evaluation*, *12* (10), 1-8.
Hu, C. (2002). Advanced tourism demand forecasting: Artificial neural network and Box-Jenkins modeling. (Ph.D. dissertation). Indiana, United States: Purdue University.
Hurley, W., & Lior, D. (2002). Combining expert judgment: On the performance of trimmed mean vote aggregation procedures in the presence of strategic voting. *European Journal of Operational Research*, *140*, 142-147.

Huth, W. L., Eppright, D. R., & Taube, P. M. (1994). The indexes of consumer sentiment and confidence: Leading or misleading guides to future buyer behavior. *Journal of Business Research*, *29* (3), 199-206.

Hyndman, R., Koehler, A., Ord, J., & Snyder, R. (2008). *Forecasting with exponential smoothing: The state space approach.* Berlin: Springer.

Hyndman, R., Koehler, A., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*, 439-454.

IHS EViews. (2009a). *EViews 7 user's guide I.* Irvine, CA: Quantitative Micro Software, LLC.

IHS EViews. (2009b). *EViews 7 user's guide II*. Irvine, CA: Quantitative Micro Software, LLC.

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

Jang, S. (2004). Mitigating tourism seasonality: A quantitative approach. *Annals of Tourism Research*, *31* (4), 819-836.

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

Joutz, F., & Stekler, H. (2000). An evaluation of the predictions of the Federal Reserve. *International Journal of Forecasting*, *16* (1), 17-38.

Kahn, H. (1979). *World economic development 1979 and beyond.* London: Croom Helm.

Kahn, H., & Weiner, A. (1967). *The year 2000.* London: Macmillan.

Katsura, T., & Sheldon, P. (2008). Forecasting mobile technology use in Japanese tourism. *Information Technology & Tourism*, *10* (3), 201-214.

Kaynak, E., Bloom, J., & Leibold, M. (1994). Using the Delphi technique to predict future tourism potential. *Marketing Intelligence & Planning, 12* (7), 18-29.

Kaynak, E., & Cavlek, N. (2007). Measurement of tourism market potential of Croatia by use of Delphi qualitative research technique. *Journal of East-West Business*, *12* (4), 105-123.

Kaynak, E., & Macaulay, J. A. (1984). The Delphi technique in the measurement of tourism market potential: The case of Nova Scotia. *Tourism Management*, *5* (2), 87-101.

Kaynak, E., & Marandu, E. E. (2006). Tourism market potential analysis in Botswana: A Delphi study. *Journal of Travel Research*, *45* (2), 227-237.

Kaynak, E., & Pathak, R. (2006). Tourism market potential of small resource-based economics: The case of Fiji Islands. In M. Adams, & A. Alkhafaji, *Business Research Yearbook: Global Business Perspectives* (pp. 123-128). Beltsville, MD, USA: The International Academy of Business Disciplines.

Kelly, J., & Warnick, R. (1999). *Recreation trends and markets: The 21st centry.* Champaign, Illinois: Sagamore.

Kibedi, G. (1981). Future trends in international tourism. *Tourism Review*, *36*(1), 3-6.

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.

Klein, H. E., & Linneman, R. E. (1984). Environmental assessment: An international study of corporate practice. *Journal of Business Strategy*, *15* (1), 66-75.
Klein, H., & Newman, W. (1980). How to use SPIRE: A systematic procedure for identifying relevant environments for strategic planning. *Journal of Business Strategy*, *1* (1), 32-45.
Kollwitz, H. (2011). Evaluating cruise demand forecasting practices: A Delphi approach. In P. Gibson, A. Papathanassis, & P. Milde, *Cruise sector challenges: Making progress in an uncertain world* (pp. 39-55). Wiesbaden: Gabler Verlag.
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.
Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, *54*, 159–178.
Lapage, W. (1994). Using panels for travel and tourism research. In J. R. Ritchie, & C. Goeldner, *Tourism and hospitality research: A handbook for managers & researchers*. New York: Wiley.
Larréché, J.-C., & Moinpour, R. (1983). Managerial judgment in marketing: The concept of expertise. *Journal of Marketing Research*, *20* (2), 110-121.
Lawrence, M., & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior and Human Decision Processes*, *43* (2), 172-187.
Lawrence, M., Edmundson, R., & O'Connor, M. (1985). An examination of the accuracy of judgmental extrapolation of time series. *International Journal of Forecasting*, *1* (1), 25-35.
Lawrence, M., Edmundson, R., & O'Connor, M. (1986). The accuracy of combining judgmental and statistical forecasts. *Management Science*, *32* (12), 1521-1532.
Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *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.
Lee, C., & Kim, J. (1998). International tourism demand for the 2002 world cup Korea: A combined forecasting technique. *Pacific Tourism Review*, *2*, 157-166.
Lee, C.-K., Song, H.-J., & Mjelde, J. W. (2008). The forecasting of international Expo tourism using quantitative and qualitative techniques. *Tourism Management*, *29* (6), 1084-1098.
Lee, M. S., Elango, B., & Schnaars, S. P. (1997). The accuracy of the Conference Boards buying plans index: A comparison of judgmental vs. extrapolation forecasting methods. *International Journal of Forecasting*, *13* (1), 127-135.
Lee, W. Y., Goodwin, P., Fildes, R., Nikolopoulos, K., & Lawrence, M. (2007). Providing support for the use of analogies in demand forecasting tasks. *International Journal of Forecasting*, *23*, 377-390.
Leitner, J., & Leopold-Wildburger, U. (2011). Experiments on forecasting behavior with several sources of information- A review of the literature. *European Journal of Operational Research*, *213* (3), 459-469.

Lennon, J., & Yeoman, I. (2007). Drivers and scenarios of Scottish tourism – Shaping the future to 2015. *Tourism Recreation Research*, *32* (1), 345-367.

Lewis, C. (1982). *Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting.* London: Butterworth Scientific.

Li, G. (2009). Tourism demand modelling and forecasting: A review of literature related to greater China. *Journal of China Tourism Research*, *5* (1), 2-40.

Li, G., Song, H., & Witt, S. F. (2006). Time varying parameter and fixed parameter linear AIDS: An application to tourism demand forecasting. *International Journal of Forecasting*, *22* (1), 57-71.

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

Lim, C. (1997). Review of international tourism demand models. *Annals of Tourism Research*, *24* (4), 835-849.

Lim, C. (1999). A meta-analytic review of international tourism demand. *Journal of Travel Research*, *37* (3), 273-284.

Lim, C. (2006). A survey of tourism demand modelling practice: Issues and implications. In L. Dwyer, & P. Forsyth, *International handbook on the economics of tourism.* Cheltenham England; Northampton, MA: Edward Elgar Pub.

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, V. S. (2013). Improving forecasting accuracy by combining statistical and judgmental forecasts in tourism. *Journal of China Tourism Research, 9*(3), 325-352.

Lin, V. S., & Song, H. (2011). An examination of Delphi forecast accuracy. *Proceedings of the 31st International Symposium on Forecasting*, (p. 38). Prague,Czech Republic.

Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry*. New York: Sage.

Linstone, H. (1978). The Delphi technique. In J. Fowles, *Handbook of futures research.* London: Greenwood Place.

Linstone, H., & Turoff, M. (2002). *The Delphi method: Techniques and applications.* Retrieved August 1, 2011, from http://is.njit.edu/pubs/delphibook/

Liu, J. C. (1988). Hawaii tourism to the year 2000: A Delphi forecast. *Tourism Management*, *9*, 279-290.

Lloyd, J., La Lopa, J. M., & Braunlich, C. G. (2000). Predicting changes in Hong Kong's hotel industry given the change in sovereignty from Britain to China in 1997. *Journal of Travel Research*, *38* (4), 405-410.

Lobo, G. J. (1991). Alternative methods of combining security analysts and statistical forecasts of annual corporate earnings. *International Journal of Forecasting*, *7* (1), 57-63.

Lobo, G. J. (1992). Analysis and comparison of financial analysts, time series, and combined forecasts of annual earnings. *Journal of Business Research*, *24* (3), 269-280.

Lobo, G. J., & Nair, R. D. (1990). Combining judgmental and statistical Forecasts - An application to earnings forecasts. *Decision Sciences*, *21* (2), 446-460.

Longsfield, K. (2004). *In-depth Interviews.* Retrieved June 1, 2012, from http://www.aidsmark.org/ipc_en/pdf/manual/14_ResearchToolkit-Ch6-In-Depth-Interviews.pdf

Loo, R. (2002). The Delphi method: A powerful tool for strategic management. *Policing: An International Journal of Police Strategies & Management*, *25* (4), 762-769.

Lundtorp, S. (2001). Measuring tourism seasonality. In T. Baum, & S. Lundtorp, *Seasonality in tourism.* Oxford, UK: Elsevier Science Ltd.

Lusk, C., & Hammond, K. R. (1991). Judgment in a dynamic task: Microburst forecasting. *Journal of Behavioral Decision Making*, *4* (1), 55-73.

Lusk, C., Stewart, T. R., Hammond, K. R., & Potts, R. J. (1990). Judgment and decision making in dynamic tasks: The case of forecasting the microburst. *5* (4), 627-639.

Lyness, K. S., & Cornelius, E. T. (1982). A comparison of holistic and decomposed judgment strategies in a performance rating simulation. *Organizational Behavior and Human Performance*, *29* (1), 21-38.

MacGregor, D. G. (2001). Decomposition for judgmental forecasting and estimation. In J. Armstrong, *Principles of forecasting: A handbook for researchers and practitioners.* Dordrecht: Kluwer Academic.

MacGregor, D., Lichtenstein, S., & Slovic, P. (1988). Structuring knowledge retrieval: An analysis of decomposed quantitative judgments. *Organizational Behavior and Human Decision Processes*, *42* (3), 303-323.

MacKinnon, J. G. (1996). Numerical distribution functions for unit root and cointegration tests. *Journal of Applied Econometrics*, *11* (6), 601–618.

Makridakis, S., & Winkler, R. (1983). Averages of forecasts: Some empirical results. *Management Science*, *29* (9), 987-996.

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., et al. (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.

Manning, R., & Fraysier, M. (1989). Expert and public opinion: Conflicting or complementary views? *Journal of Park and Recreation Administration*, *7* (3), 44-59.

Marmier, F., & Cheikhrouhou, N. (2010). Structuring and integrating human knowledge in demand forecasting: A judgemental adjustment approach. *Production Planning & Control*, *21* (4), 399-412.

Martin, W., & Mason, S. (1998). *Transforming the future: Rethinking free time and work.* Sudbury: Leisure Consultants.

Martino, J. (1983). *Technological forecasting for decision making.* New York: American Elsevier.

Martino, J. (1970). The consistency of Delphi forecasts. *The Futurist*, *4*, 63-64.

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., & Diamantopoulous, A. (1990). Judgmental revision of sales forecasts: Effectiveness of forecast selection. *Journal of Forecasting*, *9* (4), 407-415.

Mathews, B. P., & Diamantopoulos, A. (1992). Judgemental revision of sales forecasts: The relative performance of judgementally revised versus non-revised forecasts. *Journal of Forecasting*, *11*, 569-576.
McCleary, K. W., & Whitney, D. L. (1994). Projecting western consumer attitudes toward travel to six eastern European countries. In M. Uysal, *Global tourist behavior* (pp. 239-256). New York: International Business Press.
McCubbrey, D. J. (1999). Disintermediation and reintermediation in the U.S. air travel distribution industry: A Delphi Study. *Communications of the Association for Information Systems*, *1* (18), 1-39.
McCubbrey, D. J., & Taylor, R. G. (2005). Disintermediation and reintermediation in the U.S. air travel distribution industry: A Dephi reprise. *Communications of the Association for Information Systems*, *2005*, 464-477.
McGregor, D. (1938). The major determinants of the prediction of social events. *Journal of Abnormal and Social Psychology*, *33* (2), 179-204.
McGregor, D., & Lichtenstein, S. (1991). Problem structuring aids for quantitative estimation. *Journal of Behavioral Decision Making*, *4* (2), 101-116.
Miller, G. (2001). The development of indicators for sustainable tourism: Results of a Delphi survey of tourism researchers. *Tourism Management*, *22*, 351-362.
Moeller, G. H., & Shafer, E. L. (1983). The use and misuse of Delphi forecasting. In S. Lieber, & D. Fesenmaier, *Recreation planning and management* (pp. 96-104). State College, Pennsylvania: Venture Publishing.
Moeller, G. H., & Shafer, E. L. (1994). The Delphi technique: A tool for long-range travel and tourism planning. In J. R. Ritchie, & C. R. Goeldner, *Travel, tourism, and hospitality research: A handbook for managers and researchers* (2 ed., pp. 473-480). New York: Wiley.
Morley, C. L. (2009). Dynamics in the specification of tourism demand models. *Tourism Economics*, *15* (1), 23-39.
Moutinho, L., & Witt, S. F. (1995). Forecasting the tourism environment using a consensus approach. *Journal of Travel Research*, *33* (4), 46-50.
Müller, H. (1998). Long-haul tourism 2005 - Delphi study. *Journal of Vacation Marketing*, *4* (2), 193-201.
Murray, T. J. (1979). Delphi methodologies: A review and critique. *Urban Systems*, *4* (2), 153-158.
Musso, A., & Phillips, S. (2002). *Comparing projections and outcomes of IMF-supported programs.* IMF Staff Papers (Vol. 49, No. 1), International Monetary Fund.
Nadal, J. R., Font, A. R., & Rossello, A. S. (2004). The economic determinants of seasonal patterns. *Annals of Tourism Research*, *31* (3), 697-711.
Narayan, P. K., & Smyth, R. (2006). Higher education, real income and real investment in China: Evidence from granger causality tests. *Education Economics*, *14* (1), 107-125.
Ng, D. (1984). A model estimating the demand for leisure services manpower. *World Leisure and Recreation*, *26* (5), 45-49.
Ng, S., & Perron, P. (2001). Lag length selection and the construction of unit root tests with good size and power. *Econometrica*, *69* (6), 1519-1554.
Nordhaus, W. D. (1987). Forecasting efficiency: Concepts and applications. *The review of economics and statistics*, *69* (4), 667-674.
O'Connor, M. (1989). Models of human behaviour and confidence in judgement: A review. *International Journal of Forecasting*, *5* (2), 159-169.

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

O'Connor, M., Remus, W., & Lim, K. (2005). Improving judgmental forecasts with judgmental bootstrapping and task feedback support. *Journal of Behavioral Decision Making*, *18* (4), 247-260.

O'Neil, S. (2009). *Basic statistics.* Retrieved October 6, 2012, from University of Pretoria: http://web.up.ac.za/sitefiles/file/40/1111/Basic%20Statistics%20for%20the%20utterly%20confused.pdf

Obermair, K. (1998). *Future trends in tourism: Alliance internationales de tourisme Delphi study.* Vienna: Alliance Internationales de Tourisme.

Ogburn, W. F. (1934). Studies in prediction and the distortion of reality. *Social Forces*, *13* (2), 224-229.

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: An example, design considerations and applications. *Information & Management*, *42* (1), 15-29.

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

Önkal, D., Thomson, M. E., & Pollock, A. C. (2007). Judgmental forecasting. In M. P. Clements, & D. F. Hendry, *A companion to economic forecasting* (pp. 133-151). Oxford: Blackwell.

Pan, S. Q., Vega, M., Vella, A. J., Archer, B. H., & Parlett, G. R. (1995). A mini-Delphi approach: An improvement on single round techniques. *Progress in Tourism and Hospitality Research*, *2*, 27-39.

Parackal, M., Goodwin, P., & O'Connor, M. (2007). Judgement in forecasting. *International Journal of Forecasting*, *23* (3), 343-345.

Parente, F., & Anderson-Parente, J. (1987). Delphi inquiry systems. In G. Wright, & P. Ayton, *Judgmental forecasting.* New York: Wiley.

Parente, R., & Anderson-Parente, J. (2011). A case study of long-term Delphi accuracy. *Technological Forecasting and Social Change*, *78*, 1705-1711.

Patterson, M., & McDonald, G. (2004). *How clean and green is New Zealand tourism? Lifecyle and future environmental impacts.* Lincoln, Canterbury, New Zealand: Manaaki Whenua Press, Landcare Research.

Pereira, B., Coqueiro, R. C., & Perrota, A. H. (1989). Experience in combining subjective and quantitative forecasts of open market rates. *Journal of Forecasting*, *8* (3), 343-348.

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

Pizam, A. (1994). Planning a tourism research investigation. In J. R. Ritchie, & C. Goeldner, *Travel, tourism, and hospitality research* (pp. 91-104). New York: Wiley.

Pollack-Johnson, B. (1995). Hybrid structures and improving forecasting and scheduling in project management. *Journal of Operations Management,*, *12* (2), 101-117.

Prideaux, B., Laws, E., & Faulkner, B. (2003). Events in Indonesia: Exploring the limits to formal tourism trends forecasting methods in complex crisis situations. *Tourism Management*, *24* (4), 475-487.

Reid, D. J. (1968). Combining three estimates of gross domestic product. *Economica*, *35* (140), 431-444.

Remus, W. (1984). An empirical investigation of the impact of graphical and tabular data presentations on decision making. *Management Science*, *30* (5), 533-542.

Remus, W. (1987). A study of graphical and tabular displays and their interaction with environmental complexity. *Management Science*, *33* (9), 1200-1204.

Remus, W., O'Connor, M., & Griggs, K. (1996). Does feedback improve the accuracy of recurrent judgmental forecasts? *Organizational Behavior and Human Decision Processes*, *66* (1), 22-30.

Remus, W., O'Connor, M., & Griggs, K. (1998). The impact of incentives on the accuracy of subjects in judgmental forecasting experiments. *International Journal of Forecasting*, *14* (4), 515-522.

Ritchie, J. B. (1985). The nominal group technique: An approach to consensus policy formulation in tourism. *Tourism Management*, *6* (2), 82-94.

Ritchie, J. B. (1994). The nominal group technique: Applications to tourism research. In J. Ritchie, *Travel, Tourism, Hospitality Research* (pp. 439-448). New York: Wiley.

Roche, M. (2000). *Mega-events and modernity: Olympics and expos in the growth of global culture.* London, UK: Routledge.

Rossetto, A. (1999). Using scenarios in forecasting. *The Council for Australian University Tourism and Hospitality Education (CAUTHE).* Adelaide: BTR Conference Paper 99.3.

Rothe, J. T. (1978). Effectiveness of sales forecasting methods. *Industrial Marketing Management*, *7* (2), 114-118.

Rowe, A., Smith, J. D., & Borein, F. (2002). *Career award in travel and tourism: Standard level.* Cambridge: Cambridge University Press.

Rowe, G. (1998). The use of structured groups to improve judgmental forecasting. In G. Wright, & P. Goodwin, *Forecasting with judgment* (pp. 201-235). Chichester, England: Wiley.

Rowe, G., & Wright, G. (2001). Expert opinions in forecasting: The role of the Delphi technique. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 125-144). Dordrecht: Kluwer Academic.

Rowe, G., & Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting*, *12*, 73-89.

Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting*, *15* (4), 353-375.

Rushdi, M., J. Kim, & Silvapulle,P. (2012). ARDL bounds tests and robust inference for the long run relationship between real stock returns and inflation in Australia. *Economic Modelling*, *29*, 535-543.

Sadi, M. A., & Henderson, J. C. (2005). Tourism in Saudi Arabia and its future development. *Journal of Business and Economics*, *11*, 94-111.

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. P. (1990). Some empirical findings on short-term forecasting: Technique complexity and combinations. *Decision Sciences*, *20* (3), 635-640.

Sanders, N., & Ritzman, L. P. (1992). The need for contextual and technical knowledge in judgemental forecasting. *Journal of Behavioral Decision Making*, *5*, 39-52.

Sanders, N., & Ritzman, L. P. (1995). Bringing judgment into combination forecasts. *Journal of Operations Management*, *13* (4), 311-321.
Sanders, N., & Ritzman, L. P. (2001). Judgmental adjustment of statistical forecasts. In J. Armstrong, *Principles of forecasting: A handbook for researchers and practitioners* (pp. 405-416). Dordrecht: Kluwer Academic.
Sanders, N., & Ritzman, L. P. (2004). Integrating judgmental and quantitative forecasts: Methodologies for pooling marketing and operations information. *International Journal of Operations & Production Management*, *24* (5-6), 514-529.
Schnaars, S. (1984). Situational factors affecting forecast accuracy. *Journal of Marketing Research*, *21* (3), 290-297.
Schnaars, S. (1986). A comparison of extrapolation models on yearly sales forecasts. *International Journal of Forecasting*, *2* (1), 71-85.
Schnaars, S. (1987). How to develop and use scenarios. *Long Range Planning*, *20* (1), 105-114.
Schwaninger, M. (1989). Trends in leisure and tourism for 2000–2010: Scenario with consequences for planners. In S. F. Witt, & L. Mouthinho, *Tourism marketing and management handbook* (pp. 599–605). Cambridge: Prentice Hall.
Seely, R., Iglarsh, H., & Edgell, D. (1980). Utilizing the Delphi technique at international conferences: A method for forecasting international tourism conditions. *Journal of Travel Research*, *1*, 30-36.
Shafer, E.L, &Moeller, G.H. (1974). Through the looking glass in environmental management,*Parks and Recreation*,*9*(2),20-23.
Shafer, E. L., Moeller, G.H., & Getty, R. (1974). Future leisure environments. *USDA Forest Service Research Paper NE-301*. Upper Dardy, PA.
Shen, S., Li, G., & Song, H. (2008). An assessment of combining tourism demand forecasts over different time horizons. *Journal of Travel Research*, *47* (2), 197-207.
Shen, S., Li, G., & Song, H. (2011). Combination forecasts of International tourism demand. *Annals of Tourism Research*, *38*(1), 72-89.
Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.
Silver, M. S. (1991). Decisional guidance for computer-based decision support. *MIS Quarterly*, *15*, 105-122.
Singapore Tourism Board. (2011). *Monthly tourism focus*. Retrieved September 10, 2011, from Tourism statistics publications: https://app.stb.gov.sg/asp/tou/tou03.asp
Smeral, E., & Weber, A. (2000). Forecasting international tourism trends to 2010. *Annals of Tourism Research*, *27* (4), 982-1006.
Smeral, E., Witt, S. F., & Witt, C. A. (1992). Econometric forecasts: Tourism trends to 2000. *Annals of Tourism Research*, *19* (3), 450-466.
Smith, S. (1995). Forecasting tourism demand and market trends. In S. Smith, *Tourism analysis: A handbook*. Harlow, England: Longman.
Song, H., & Guo, W. (2008). Tourism demand modeling and forecasting. In A. Woodside, D. Martin, & D. Martin, *Tourism management: Analysis, behaviour and strategy* (pp. 113-128). Wallingford, UK: CABI International.
Song, H., & Hyndman, R. J. (2011). Tourism forecasting: An introduction. *International Journal of Forecasting*, *27* (3), 817-821.
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., Li, G., Witt, S. F., & Fei, B. (2010). Tourism demand modelling and forecasting: How should demand be measured? *Tourism Economics*, *16* (1), 63-81.
Song, H., Lin, S., Witt, S. F., & 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. (2000). *Tourism demand modeling and forecasting – Modern econometric approaches*. Oxford: Pergamon.
Song, H., Witt, S. F., & Jensen, T. C. (2003). Tourism forecasting: Accuracy of alternative econometric models. *International Journal of Forecasting*, *19* (1), 123-141.
Song, H., Witt, S. F., & Li, G. (2009). *The advanced econometrics of tourism demand.* New York: Routledge.
Song, H., Witt, S. F., & Lin, S. (2010). Forecasting the demand for Hong Kong tourism: The views of experts. *The 20th Annual Conference of Council for Australian University Tourism and Hospitality Education (CAUTHE) 2010:Tourism and Hospitality: Challenge the Limits*, (pp. 1397-1420). School of Management,University of Tasmania. Tasmania,Australia.
Song, H., Witt, S. F., & Zhang, X. (2008). Developing a Web-based tourism demand forecasting system. *Tourism Economics*, *14* (3), 445-468.
Song, H., Witt, S. F., Wong, K. K., & Wu, D. (2009). An empirical study of forecast combination in tourism. *Journal of Hospitality & Tourism Research*, *33*, 3-29.
Song, H., Wong, K. K., & Chon, K. K. (2003). Modelling and forecasting the demand for Hong Kong tourism. *International Journal of Hospitality Management*, *22* (4), 435-451.
Spenceley, A. (2008). Requirements for sustainable nature-based tourism in transfrontier conservation areas: A Southern African Delphi consultation. *Tourism Geographies*, *10* (3), 285-311.
Spradley, J. (1979). *The ethnographic interview.* New York: Holt, Rinehart and Winston.
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
Stewart, T. R. (2001). Improving reliability of judgmental forecasts. In J. S. Armstrong, & J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 81-106). Dordrecht: Kluwer Academic.
Stewart, T. R., & Lusk, C. M. (1994). Seven components of judgmental forecasting skill: Implications for research and the improvement of forecasts. *Journal of Forecasting*, *13*, 579-599.

Swarbrooke, J., & Horner, S. (2002). *Business travel and tourism*. Oxford: Butterworth-Heinemann.
Syntetos, A. A., Nikolopoulos, K., Boylan, J. E., Fildes, R., & Goodwin, P. (2009). The effects of integrating management judgement into intermittent demand forecasts. *International Journal of Production Economics*, *118* (1), 72-81.
Taylor, R. E., & Judd, L. L. (1994). Delphi forecasting. In S. F. Witt, & L. Moutinho, *Tourism marketing and management handbook* (2nd ed., pp. 535-539). Hertfordshire: Prentice Hall.
Tesar, G., Edgell, D., & Seely, R. (1979). Use of modified scenario research in forecasting of tourism in the U.S. *Travel Research Journal*, *1*, 49-57.
The Government of the Hong Kong Special Administrative Region. (2008). *Half-yearly economic report 2008*. Financial Secretary's Office, Economic Analysis Division, Economic Analysis and Business Facilitation Unit, Hong Kong.
The Government of the Hong Kong Special Administrative Region. (2012). *2011 Economic background and 2012 prospects*. Retrieved September 20, 2012, from Hong Kong economy: http://www.hkeconomy.gov.hk/en/pdf/er_11q4.pdf
The Legislative Council Commission. (2012). *Paper on the work plan of Hong Kong Tourism Board prepared by the Legislative Council Secretariat (Background brief)*. Retrieved September 22, 2012, from Panel on Economic Development (Papers): http://legco.gov.hk/yr11-12/english/panels/edev/papers/edev0117cb1-808-4-e.pdf
The Wallace Foundation. (2012). *Workbook E: Conducting in-depth interviews*. Retrieved July 25, 2012, from The Wallace Foundation: http://www.wallacefoundation.org/knowledge-center/after-school/collecting-and-using-data/Documents/Workbook-E-Indepth-Interviews.pdf
Tichy, G. (2004). The over-optimism among experts in assessment and foresight. *Technological Forecasting and Social Change*, *71* (4), 341-363.
Tideswell, C., Mules, T., & Faulkner, B. (2001). An integrative approach to tourism forecasting: A glance in the rearview mirror. *Journal of Travel Research*, *40* (2), 162-171.
Tisdell, C. (2000). *The economics of tourism. (Vol. 1)*. Cheltenham: Edward Elga.
Tolley, R., Lumsdon, L., & Bickerstaff, K. (2001). The future of walking in Europe: A Delphi project to identify expert opinion on future walking scenarios. *Transport Policy*, *8* (4), 307-315.
Tolley, R., Lumsdon, L., & Bickerstaff, K. (2010). The future of walking in Europe: Revisiting expert opinion ten years later. *PQN Final Report – Part B3: Documentation – The future of walking*. (D. Sauter, D. Bazik, K. Bickerstaff, E. Drápela, L. Lumsdon, L. Martincigh, et al., Eds.) Office, COST. Retrieved September 15, 2011, from Pedestrians Quality Needs (PQN): http://www.walkeurope.org/final_report/default.asp
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
Tsamakos, A., Giaglis, G. M., & Kourouthanassis, P. (2002). *Auctioning tourism products over mobile devices*. Retrieved June 25, 2010, from http://fama2.us.es:8080/turismo/turismonet1/economia%20del%20turismo/marketing%20turistico/AUCTIONING%20TOURISM%20PRODUCTS%20OVER%20MOBILE%20DEVICES.PDF
Turner, L. W., Kulendran, N., & Pergat, V. (1995). Forecasting New Zealand tourism demand with disaggregated data. *Tourism Economics*, *1* (1), 50-69.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185* (4157), 1124-1131.
Tversky, A., & Kahneman, D. (2002). Extensional versus intuitive reasoing. In T. Gilovich, D. Griffin, & D. Kahneman, *Heuristics and biases: The psychology of intuitive judgment* (pp. 19-48). Cambridge: Cambridge University Press.
UNWTO. (2001). *Tourism 2020 vision Vol. 3 East Asia & Pacific (English Version).* Retrieved June 21, 2010, from World Tourism Organization: http://www.wtoelibrary.org/content/j452p5/?p=7857e2de615b41a6a50359bf6e8844d4&pi=4
UNWTO. (2011). *Tourism towards 2030: Global overview*. Retrieved August 20, 2012, from World Tourism Organization: http://www.e-unwto.org/content/w45127/fulltext.pdf
UNWTO. (2012a). *UNWTO world tourism barometer and statistical annex (September 2012).* Retrieved September 25, 2012, from World Tourism Organization (e-unwto): http://www.e-unwto.org/content/q033422150123345/fulltext.pdf
UNWTO. (2012b). *UNWTO tourism highlights (2012 Edition).* Retrieved September 23, 2012, from World Tourism Organization: http://www.e-unwto.org/content/x6k11g/fulltext.pdf
UNWTO and ETC. (2011). *Handbook on tourism forecasting methodologies.* Madrid: World Tourism Organization and the European Travel Commission.
USTA. (2011). *The 2010 travelhorizons™ research program*. Retrieved November 21, 2011, from US Travel Assocition (USTA): http://www.ustravel.org/research/domestic-research/travelhorizons.
Uysal, M., & Crompton, J. L. (1985). An overview of approaches used to forecast tourism demand. *Journal of Travel Research*, *23* (4), 7-15.
van Doorn, J. W. (1982). Can futures research contribute to tourism policy? *Tourism Management*, *3* (3), 149-166.
van Doorn, J. W. (1984). Tourism forecasting and the policymaker: Criteria of usefulness. *Tourism Management*, *5* (2), 24-39.
van Doorn, J. W. (1986). Scenario writing: A method for long-term tourism forecasting? *Tourism Management*, *7* (1), 33-49.
Vanegas, S. M., & Croes, R. R. (2000). Evaluation of demand: US tourists to Aruba. *Annals of Tourism Research*, *27* (4), 946-963.
Vanhove, N. (2011). *The economics of tourism destinations.* London: Elsevier.
Var, T. (1984). Delphi and GSV techniques in tourism forecasting and policy design. *Problems of Tourism*, *3*, 41-52.
Varum, C. A., Melo, C., Alvarenga, A., & de Carvalho, P. (2011). Scenarios and possible futures for hospitality and tourism. *Foresight*, *13* (1), 19-35.
Vaugeois, N., Rollins, R., Hammer, K., Randall, C., Clark, B., & Balmer, K. (2005). Realizing improved knowledge transfer in the leisure industry: A new agenda and strategies to researchers and professionals. *Travel and Tourism Research Association Canada Chapter Conference Proceedings.* Kelowna, BC, Canada.
Veal, A. (2010). *Leisure, sport and tourism: Politics, policy and planning.* Wallingford, Oxfordshire: CAB International.
Wagenaar, W., & Sagaria, S. (1975). Misperception of exponential growth. *Perception and Psychophysics*, *18*, 416-422.
Walker, K. B., & McClelland, L. A. (1991). Management forecasts and statistical prediction model forecasts in corporate budgeting. *Journal of Accounting Research*, *29* (2), 371-381.

Webby, R., & O'Connor, M. (1996). Judgemental and statistical time series forecasting: A review of the literature. *International Journal of Forecasting*, *12* (1), 91-118.
Webby, R., & O'Connor, M. (2001). Judgmental time-series forecasting using domain knowledge. In J. S. Armstrong, *Principles of forecasting: A handbook for researchers and practitioners* (pp. 389-404). Dordrecht: Kluwer Academic.
Webby, R., O'Connor, M., & Edmundson, B. (2005). Forecasting support systems for the incorporation of event information: An empirical investigation. *International Journal of Forecasting*, *21* (3), 411-423.
Weber, K., & Ladkin, A. (2003). The convention industry in Australia and the United Kingdom: Key issues and competitive forces. *Journal of Travel Research*, *42* (2), 125-132.
Weinberg, C. B. (1986). Arts plan: Implementation, evolution, and usage. *Marketing Science*, *5* (2), 143-158.
Wheeller, B., Hart, T., & Whysall, P. (1990). Application of the Delphi technique : A reply to Green, Hunter and Moore. *Tourism Management*, *11* (2), 121-122.
Whyte, D. N. (1992). Key trends and issues impacting local government recreation and park administration in the 1990s. *Thesis (PhD.)*. Indiana University.
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.
Winklhofer, H. M., & Diamantopoulos, A. (2002). Managerial evaluation of sales forecasting effectiveness: A MIMIC modeling approach. *International Journal of Research in Marketing*, *19* (2), 151-166.
Winklhofer, H. M., Diamantopoulos, A., & Witt, S. F. (1996). Forecasting practice: A review of the empirical literature and an agenda for future research. *International Journal of Forecasting*, *12* (2), 193-221.
Witt, S. F., & Moutinho, L. (1989). *Tourism marketing and management handbook.* Hertfordshire: Prentice Hall.
Witt, S. F., & Witt, C. A. (1995). Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting*, *11* (3), 447-475.
Witt, S. F., & Witt, C. A. (1992). *Modeling and forecasting demand in tourism.* London: Academic Press.
Wolfe, C., & Flores, B. (1990). Judgmental adjustment of earnings forecasts. *Journal of Forecasting*, *9* (4), 389-405.
Wong, K., Song, H., Witt, S. F., & Wu, D. C. (2007). Tourism forecasting: To combine or not to combine? *Tourism Management*, *28* (4), 1068-1078.
Woudenberg, F. (1991). An evaluation of Delphi. *Technological Forecasting and Social Change*, *40* (2), 131-150.
Wright, G., & Ayton, P. (1987). *Judgmental forecasting.* New York: Wiley.
Wright, G., & Goodwin, P. (1998). *Forecasting with judgment.* Chichester: Wiley.
Wright, G., Lawrence, M. J., & Collopy, F. (1996). The role and validity of judgment in forecasting. *International Journal of Forecasting*, *12* (1), 1-8.
WTTC. (2011). *Travel & tourism 2011.* Retrieved September 20, 2012, from World Travel & Tourism Council: http://www.wttc.org/site_media/uploads/downloads/traveltourism2011.pdf
WTTC. (2012). *Economic data search tool.* Retrieved September 20, 2012, from World Travel & Tourism Council: http://www.wttc.org/research/economic-data-search-tool/

Yacoumis, J. (1980). Tackling seasonality: The case of Sri Lanka. *International Journal of Tourism Management*, *1* (2), 84-98.
Yang, Y.(2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika Trust*, *92*(4),937-950.
Yaniv, I., & Hogarth, R. M. (1993). Judgmental versus statistical prediction: Information asymmetry and combination rules. *Psychological Science*, *4* (1), 58-62.
Yeoman, I. (2008). *Tomorrow's tourist: Scenarios & trends.* Oxford: Elsevier Science.
Yeoman, I., & Lederer, P. (2005). Scottish tourism: Scenarios and vision. *Journal of Vacation Marketing*, *11*, 71-87.
Yeoman, I., Lennon, J. J., & Black, L. (2005). Foot-and-mouth disease: A scenario of reoccurrence for Scotland's tourism industry. *Journal of Vacation Marketing*, *11*, 179-190.
Yeong, Y. W., Keng, K. A., & Leng, T. L. (1989). A Delphi forecast for the Singapore tourism industry: Future scenario and marketing implications. *European Journal of Marketing*, *23* (11), 15-26.
Yokum, T., & Armstrong, J. S. (1995). Beyond accuracy: Comparison of criteria used to select forecasting methods. *International Journal of Forecasting*, *11*, 591-597.
Zhang, H. Q., Kulendran, N., & Song, H. (2010). Measuring returns on Hong Kong's tourism marketing expenditure. *Tourism Economics*, *16* (4), 853-865.
Zivot, E., & Wang, J. (2006). Unit root tests. In E. Zivot, & J. Wang, *Modeling financial time series with S-PLUS* (2nd ed., pp. 111-139). Springer.