Pergamon

PII: S0261-5177(98)00053-3

Tourism Management, Vol. 19, No. 6, pp. 515–520, 1998

Printed in Great Britain
0261-5177/98/$19.00+0.00

Forecasting tourism: a combined approach

Fong-Lin Chu

Graduate Institute of San Min Chu I, National Taiwan University, No. 1, Roosevelt Road, Sec. 4, Taipei, Taiwan

In this article, we employ a combined seasonal nonseasonal ARIMA and sine wave nonlinear regression forecast model to predict international tourism arrivals, as represented by the number of world-wide visitors to Singapore. Compared with a similar study of the accuracy of international tourist arrivals forecasts by Chan (*Journal of Travel Research*, 1993, 31, 58–60)[1] and Chu (*Journal of Travel Research*, 1998, 36, 79–84)[2] using other univariate time series models, our proposed model has the smallest mean absolute percentage error. © **1998 Elsevier Science Ltd. All rights reserved.**

Keywords: combined forecast model, seasonal nonseasonal ARIMA, sine wave nonlinear regression, Singapore, mean absolute percentage error

Tourism has become one of the largest and most rapidly growing sectors in the world economy during the second half of the twentieth century.[3] Perhaps more than any other single activity, tourism characterizes aspects of the 'post-industrial society' and presents insights into major trends for the future. In the 30-year period since the 1950s toward the end of the 1980s, total international tourist flows have grown by a factor of six, to approximately 400 million. The impact of this growth, in terms of employment, balance of payments, the global economy generally, and the fostering of international understanding, has been considerable.[4] Since the tourist industry contributes a significant share of the gross domestic product and earns an impressive amount of foreign exchange, tourism planning is vital. In the process of tourism planning, the forecast of tourist volume in the form of arrivals is especially important since it is an indicator of the demand that can provide basic information for planning and policymaking. Tourist flows have a bearing on basic elements, such as the occupancy rate of hotels, investment in transportation and accommodation, the souvenir industry, and promotion and information.[5] Therefore, both the private sector and government bodies can use the basic data to plan their future operations and to foresee the need for facilities and infrastructure development.

A variety of forecasting techniques[1-4] is already available as the decades have progressed. Various attributes, such as accuracy of the forecasts generated, ease of use of the forecasting technique, cost of producing the forecasts, and the speed with which the forecasts can be produced, are considered when choosing from among these techniques. Notwithstanding the above, accuracy is the most important characteristic of a forecast.

Forecasting techniques may be split into quantitative and qualitative approaches. Makridakis and Hibon[6] found that the quantitative techniques give more accurate forecasts than judgment forecasts, such as those based on opinions of the sales staff and corporate executives. Quantitative methods are further divided into causal models and the time series approach. The causal model approach involves identifying functional relationships between variables under consideration and thus may depend on some predictor variables such as price index, gross domestic product, and exchange rate which are themselves not available at the moment of prediction. We therefore have to estimate all these variables, and the accuracy of our forecasts will then depend upon the precision of the estimates of other variables. This is one disadvantage of using the causal model approach. In time series model, past history on the forecast variable is merely extrapolated. There are two aspects to the study of time series-analysis and modeling. The aim of analysis is to summarize the properties of a series and to characterize its salient features. The distinguishing feature of a time series model, as opposed to a causal model, is that no attempt is made to relate a variable, say, X_t to other variables. The movements in X_t are explained solely in terms of its own past, or by its position in relation to time.[7] As far as forecast

is concerned, time series models are generally cheaper than causal models, and may also be used where causal models are inappropriate because of lack of data or incomplete knowledge regarding the causal structure. Furthermore, trend extrapolation models often give better predictions than causal models in the short run.[8]

The objective of this study is to examine the accuracy of a combination of time series forecasting models in predicting international tourism arrivals, as represented by the number of world-wide visitors to Singapore. Previous studies in the forecasting tourism literature have not considered the combination of forecasts in projecting the volume of international tourism. This paper serves to fill this gap. A combined seasonal nonseasonal ARIMA and sine wave nonlinear regression model is used to forecast the volume of tourist arrivals from January 1989 to July 1990. The results are then compared with a similar study of the accuracy on international tourist arrivals forecasts by Chan[1] and Chu.[2] The results show that our proposed model has the smallest mean absolute percentage of error.

Model specification

There seems to be general agreement among knowledgeable professionals that the typical tourism forecast cannot generally be thought of as being in any sense optimal. The consequence has been that a forecast may well be improved. One way to achieve this would be to consider two or more forecasts of the same quantity since it is often the case in tourism projection that competing forecasts are available. Suppose one has several forecasts, then it is quite possible that any one of them contains useful information absent in the others, and so rather than discard all but one forecast, it might well be profitable to incorporate them all into an overall combined forecast. A particularly simple way to achieve this is to let the combined forecast be a weighted average, with appropriately chosen weights, of the individual forecasts.

Consider a case of combining two one-step ahead forecasts.[9] Let $f_n^{(1)}$ and $f_n^{(2)}$ be forecasts of X_n using different approaches, with resulting errors

$$e_n^{(j)} = X_n - f_n^{(j)} \quad j = 1,2 \tag{1}$$

such that

$$E(e_n^{(j)}) = 0,\ E(e_n^{(j)2}) = \sigma_j^2 \quad j = 1,2$$

and

$$E(e_n^{(1)} e_n^{(2)}) = \rho\sigma_1\sigma_2$$

Consider now a combined forecast, taken to be a weighted average of two individual forecasts,

$$C_n = kf_n^{(1)} + (1-k)f_n^{(2)} \tag{2}$$

The weights k are restricted to be $0 < k < 1$, which implies the belief that both forecasts have something positive to contribute to the contribution. If the constituent forecasts have been prepared at all carefully and are based on some relevant data, nonnegative weights seem appropriate.

The forecast error is

$$e_n^{(c)} = X_n - C_n = ke_n^{(1)} + (1-k)e_n^{(2)} \tag{3}$$

Hence the error variance is

$$\sigma_c^2 = k^2\sigma_1^2 + (1-k)^2\sigma_2^2 + 2k(1-k)\rho\sigma_1\sigma_2 \tag{4}$$

This expression is minimized for the value of k given by

$$k_0 = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{5}$$

and substituting into eqn (4) yields the minimum achievable error variance as

$$\sigma_{c,0}^2 = \frac{\sigma_1^2\sigma_2^2 - (1-\rho^2)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{6}$$

Note $\sigma_{c,0}^2 < \min(\sigma_1^2, \sigma_2^2)$ unless either ρ is equal to σ_1/σ_2 or to σ_2/σ_1. If either equality holds, then the variance of the combined forecast is equal to the smaller of the two error variances. Thus, in theory, a combined forecast will usually be superior to either constituents. This means that one should never lose by combining and will usually gain. Unfortunately, a formula such as (5) cannot be used to find the combination for k the elements of variances and covariances, σ_1^2, σ_2^2 and ρ, are never known. They can of course be estimated from past forecast errors if these are available. However, it is not clear how many past errors should be used, as the forecasters who are producing the forecasts being combined may be changing in ability, at least relatively. For example, $f_n^{(1)}$ may come from a simple regression model which is applied automatically, whereas the forecasts $f_n^{(2)}$ could come from an econometrician who is continually updating and improving his model as well as applying subjective corrections and learning from past mistakes. It would be reasonable to expect that the econometrician's forecasts should be given increasing weights through time as his forecasting ability improves relative to $f_n^{(1)}$. Another problem with applying eqn (5) directly is that often only a few previous forecast errors are available, and so it is very difficult to get a satisfactory estimate for the covariance term. To circumvent these problems, Bates and Granger[9] considered a number of alternative choices of weights, including the following:

$$k_n = \frac{\sum_{t=n-\eta}^{n-1} e_t^{(2)2}}{\sum_{t=n-\eta}^{n-1} (e_t^{(1)2} + e_t^{(2)2})} \tag{7}$$

where $e_t^{(j)}$, $j = 1,2$ are sample errors generated from the forecasts, $f_n^{(1)}$ and $f_n^{(2)}$ respectively. An appropriate value for η would be 12, although this is a

rather arbitrary choice, and if fewer than 12 previous errors are available, the sums are just over those errors that are known.

Since our study deals with forecasting tourist arrivals, and our focus is on monthly data which has the element of seasonality, the choice of $e_t^{(j)}$, $j = 1,2$ becomes very important. Here we propose a way of computing k_n by using the latest lagged forecast error. In general, information about the forecast errors, e_t^1 and e_t^2 could aid immensely in determining the time-varying weight k_n in eqn (7). Although e_t^1 and e_t^2 are not known at time t when the combining weights must be calculated, e_{t-12}^1 and e_{t-12}^2 (we use $t-12$ instead of $t-1$ because of monthly data of a seasonal nature), which are known at time t, may provide some information about e_t^1 and e_t^2. Similarly e_{t-24}^1 and e_{t-24}^2 may provide some information about e_t^1 and e_t^2 but such information is less revealing than that of e_{t-12}^1 and e_{t-12}^2. We believe for monthly data with a seasonal pattern, the more recent data should reflect the trend of the series better than the less recent data. In this study, for ease of computation, the available latest lagged forecast error was considered in finding k_n. With this in mind, eqn (7) becomes

$$k_n = \frac{e_t^{(2)2}}{e_t^{(1)2} + e_t^{(2)2}}, \quad t = n - 12 \tag{8}$$

As an example, eqn (8) indicates that, to find k_n for January 1989, the relevant $e_t^{(j)}$, $j = 1,2$ are the forecast errors obtained from January 1988. With monthly seasonal data, the January 1988 forecast error is the best reference point for predicting arrivals in January 1989. Similarly, the weight k_n for February 1989 would be calculated by employing $e_t^{(j)}$, $j = 1,2$ from February 1988. At the time of finding k_n for January 1989, if errors from January 1988 are, for some reason, not yet available, we will utilize errors generated in January 1987. We believe this is an appropriate way to cope with monthly data with the pattern of seasonality.

Findings

The data[13] of international tourist arrivals are plotted to examine the pattern of international tourist arrivals in Singapore (*Figure 1*). It was found that the plot exhibited a permanent deterministic pattern of long term upward trend with short term fluctuations that were independent from one time period to the next. Note also that Singapore's popularity has been growing; the series appear to be nonstationary in that the mean is increasing over time.

We first use data from July 1977 to December 1987 to fit the ARIMA model and sine wave time series regression model. The former is done by *Box-Jenkins analysis*,[10] which involves three steps: identification, estimation, and diagnostic checking. A typical seasonal–nonseasonal ARIMA model has the form

$$\Gamma_p(B)\gamma_P(B^s)\nabla^d\nabla^D{}_sX_t = \theta_Q(B^s)\Theta_q(B)\xi_t \tag{9}$$

where $\Gamma_p(B)$ is the nonseasonal autoregressive (AR) operator, $\Theta_q(B)$ is the nonseasonal moving average (MA) operator, $\gamma_P(B^s)$is the seasonal AR operator, $\theta_Q(B^s)$is the seasonal MA operator, and $\nabla^d\ \nabla^D{}_s$ are differencing operators. In eqn (9) X_t has both seasonal and nonseasonal components, and is differ-

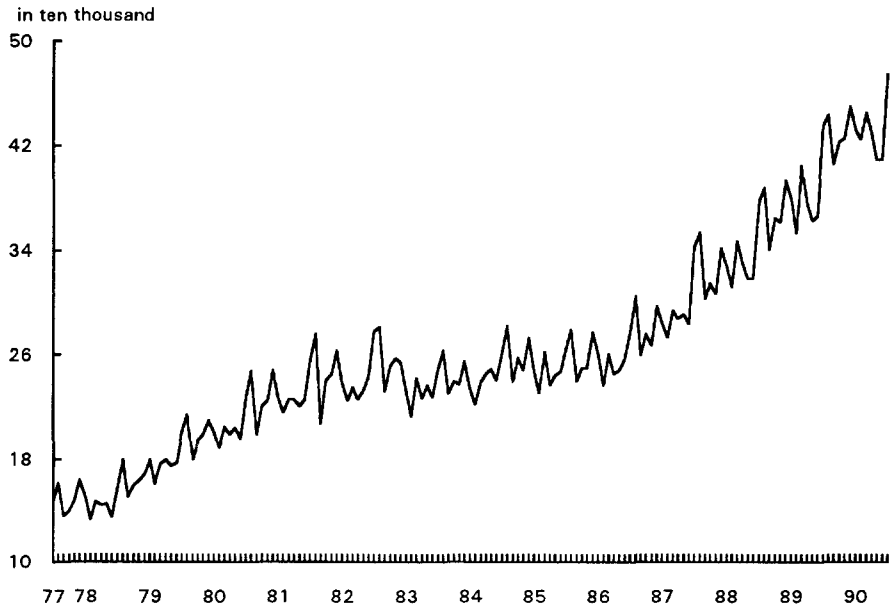

Figure 1 Monthly tourist arrivals

enced d times (length one) and D times (length s). Furthermore, the seasonal and nonseasonal AR elements are multiplied by each other as are the seasonal and nonseasonal MA elements. Equation (9) is referred to as an ARIMA $(p,d,q,)(P,D,Q,)$ process. The lower-case letters (p,d,q) indicate the nonseasonal orders and the upper-case letters (P,D,Q) denote the seasonal orders of the process. The parentheses mean that the seasonal and nonseasonal elements are multiplied by each other. Writing eqn (9) explicitly is rather tedious. Suppose, as an example, we have $p = d = q = P = D = Q = 1$. With monthly data, $s = 12$, and thus eqn (9) becomes a multiplicative ARIMA$(1,1,1)(1,1,1)_{12}$ written explicitly as

$$X_t = \Gamma_1 X_{t-1} + \gamma_{12} X_{t-12} - \Gamma_1 \gamma_{12} X_{t-13} + \xi_1 - \Theta_1 \xi_{t-1} - \theta_{12} \xi_{t-12} + \Theta_1 \theta_{12} \xi_{t-13}$$

X_t obviously depends on its past values, current error and past random errors only. The problem is to find values for $(p,d,q)(P,D,Q)$ and estimates for the parameters $\gamma_i, \theta_i, \Gamma_i$ and Θ_i such that the model 'fits' the data as closely as possible. Our exercise, carefully following three steps in Box-Jenkins analysis,[10] eventually indicated that ARIMA$(3,1,8)(1,0,0)_{12}$ is the best model. The three steps are: identification, estimation, and diagnostic checking. These steps offered criteria used to select the best fitted model. We therefore use it to predict tourist arrivals during January 1988 and December 1988.

A sine wave time series regression model[1] has the functional form:

$$X_t = \varphi_1 + \varphi_2 t + \varphi_3 \sin(\varphi_4 + \varphi_5 t) + \varepsilon_t \qquad (10)$$

where $X_t, t, \varphi_1, \varphi_2, \varphi_3, \varphi_4, \varphi_5$, and ε_t are the seasonally adjusted number of tourist arrivals at time t, time in month with respect to a fixed reference point, intercept of the linear model, slope of the linear model, amplitude of the sine function, phase angle of the sine function, frequency of the sine function and error term which is normal, independently distributed. The first two terms $\varphi_1 + \varphi_2 t$ reflect the linear trend of the data. The term $\varphi_3 \sin(\varphi_4 + \varphi_5 t)$ explains the periodic trend of the data, which can be interpreted as the magnitude of the deviation from the simple linear model at time t. Before the model is fitted to the raw data, they are adjusted for seasonal variations as the numbers of tourist arrivals go up in the vacation months such as July, August and December. Using the method of ratio to moving average, 12 monthly seasonal indices for the data are obtained and used to adjust the data and to modify forecasts. A nonlinear procedure (nonlinear least square in TSP package) is used to fit the deseasonalized data.

Data from July 1977 to December 1987 are used to forecast tourist arrivals in Singapore for the following 12 months (January 1988 to December 1988), in an effort to track the forecast errors from these months. These errors, in turn, are employed to calculate k_n in eqn (8). Then eqn (2) is used to obtain the combined ARIMA and sine wave regression forecast. *Table 1* displays the results of our forecasts using the calculated values of k_n and the forecast values obtained by Chan[1] and Chu.[2] The forecast period is from January 1989 to July 1990. We chose this period in order to compare our results with that of Chan and Chu as both implemented the forecast for the same time span.The maximum error among the forecasts occurs in May 1990, which is about 3.71% of the actual value, while the minimum error occurred in August 1989 and January 1990. The errors are about 0.08% of the actual values. The forecast of the total number of tourist arrivals in 1989 is 4824116, while the actual number of arrivals in that year was 4829950. The model underpredicted the arrivals by 5834, which is about 0.12% of the actual number. The forecast of the total number of arrivals between January 1990 and July 1990 is 3084948, while the actual number of arrivals in that period was 3030099. The error of 54849 is about 1.8% of the actual number of arrivals. For the whole forecast period, the forecast and actual number of arrivals are 7909114 and 7860049 respectively. The forecast error was 49065, which is approximately 0.62% of the actual number of arrivals. *Figure 2* shows the actual and forecast number of arrivals. The forecast follows the overall trend in tourist arrivals, and it does capture the cyclical fluctuations that occurred during this period. Moreover, the forecasts reproduce most of the turning points.

The mean absolute percentage error, MAPE, is a useful measure for comparing the accuracy of

Table 1 Forecasts of tourist arrivals between January 1989 and July 1990

Year/month	Forecast	Actual	Abs. error	% Error
1989				
Jan	378862	380024	1162	0.31
Feb	363683	353852	9831	2.78
Mar	398368	404794	6426	1.59
Apr	384103	375673	8430	2.24
May	373638	363122	10516	2.90
Jun	372092	367298	4794	1.31
Jul	431177	435974	4797	1.10
Aug	443750	443399	351	0.08
Sep	394894	406419	11525	2.84
Oct	419070	423403	4333	1.02
Nov	415952	425947	9995	2.35
Dec	448577	450045	1468	0.33
1990				
Jan	433915	433549	366	0.08
Feb	418886	425093	6207	1.46
Mar	453865	445492	8373	1.88
Apr	440031	430812	9219	2.14
May	425568	410356	15212	3.71
Jun	424669	409917	14752	3.60
Jul	488014	474880	13134	2.77

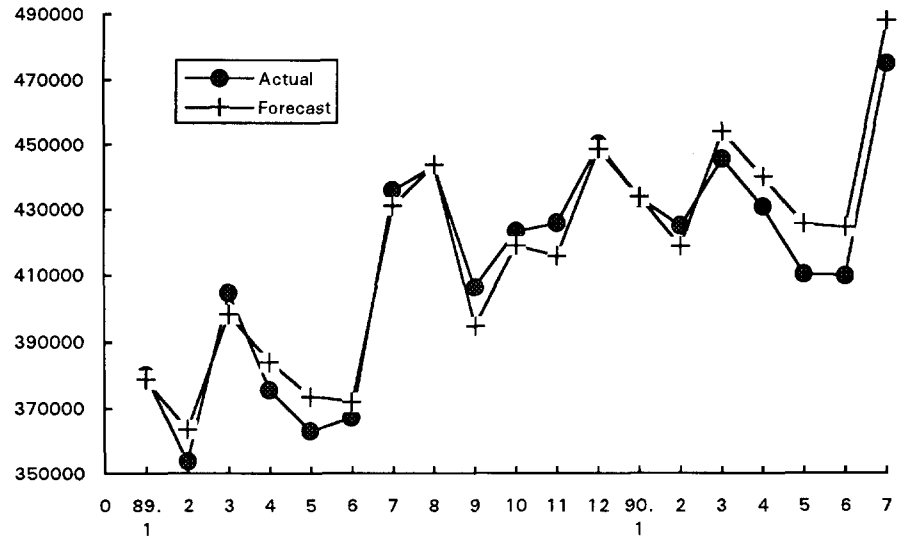

Figure 2 Actual and forecast number of tourist arrivals

forecasts between different forecasting models since it measures *relative* performance. It is also useful for conveying the accuracy of a model to managers or other nontechnical users. If an error is divided by the corresponding observed value, we have a percentage error. The MAPE is simply the mean of the absolute values of these percentage errors:

$$\frac{100}{n}\sum_{t=1}^{n}\left|\frac{(ERROR)_t}{X_t}\right| \qquad (11)$$

where the summation is across all n available absolute percentage errors. Applying eqn (11) to the errors for our proposed model gives a MAPE of 1.813%, as shown in *Table 2*.

Chan's[1] study of forecasting models on arrivals at Singapore revealed that the sine wave nonlinear regression outperformed simple linear regression, ARIMA(2,1,2), Naive I, and Naive II. Here Naive I forecast is defined as the forecast for period $t+1$ is equal to the actual number of visits in period t, and Naive II, the forecast for period $t+1$ is equal to the actual number of visits in period t multiplied by the growth rate over the previous period. The MAPE of these models are shown again in *Table 2*. Chu[2] argued that, in terms of forecasting, ARIMA $(3,1,0)0,1,0)_{12}$ did a better job than the sine wave nonlinear regression model, as the MAPE improved in *Table 2*. The combined ARIMA and sine wave nonlinear regression forecasts that we proposed performed better than the ARIMA$(3,1,0)0,1,0)_{12}$ model, and certainly better than the sine wave nonlinear regression model. In terms of MAPE, our proposed model improved 2.4% over ARIMA$(3,1,0)0,1,0)_{12}$ and 42% over the sine wave nonlinear regression model. Keep in mind that ARIMA models are very hard to beat, especially when it comes to dealing with short term forecasting. Thus a model such as the one we proposed is superior since it outperformed the ARIMA model. When compared with the Naive I model, which requires only the latest observation, our model improved approximately 205% over it. Apparently, the gain in accuracy is worth the effort although our proposed model needs a reasonable amount of data to obtain the estimated model for forecasting.

Table 2 MAPE of various models

Model	MAPE
Simple linear regression model	17.567%*
ARIMA(2,1,2)	8.577*
Naive I model	5.526*
Naive II model	4.459*
Sine wave nonlinear model	2.567*
ARIMA$(3,1,0)(0,1,0)_{12}$	1.857†
Combined forecast model	1.813

Sources: * and † are Chan[1] and Chu,[2] respectively.

We use the combination of the ARIMA model with the Sine Wave regression model because empirical studies have showed thus far that the ARIMA model and Sine Wave regression are two of the best time series models (ARIMA is in fact better than Sine Wave regression), as far as Singapore and this forecasting period (January 1989 to July 1990) are concerned. We therefore attempt to combine these best models and see if our new model can in some way further improve our forecast performance. Certainly, one can combine ARIMA$(3,1,0)(0,1,0)_{12}$ (MAPE = 1.857%) with Naive I model (MAPE = 5.526%) and/or with the simple linear regression model (MAPE = 17.567%),

but we doubt that the outcome of MAPE from such combined models will outperform that of our combination.

Summary

Singapore is one of the Asian-Pacific countries and tourism has always been an important part of the economy. According to the latest statistics, 7.3 million visitors in 1996 visit this 14 mile wide 25 mile long island. This figure outnumbered its population of 2.7 million by more than twice.[11] World Tourism Organization statistics[15] indicated Singapore was one the top destinations in 1990, ranking 12th in the world and second in Asia in terms of tourism earnings. Between 1985 and 1990 tourism was one of the fastest growing industries, with visitor arrivals growing at an average rate of 12% annually. Tourism earnings increased at more than 21% a year during the same period. With such a remarkable performance in the tourism industry, forecasting plays an essential role in tourism planning process. Because of the perishable nature of the product (for instance, unfilled airline seats and unused hotel rooms cannot be stockpiled) in the tourism industry, the need to forecast *accurately* is especially acute in the planning process.[12].

Ex post forecasts up to 19 months ahead were made for the destination country, Singapore, using mean absolute percentage error as the evaluation criterion. This article has shown that a combined ARIMA and sine wave nonlinear regression forecasting approach outperformed the ARIMA model proposed by Chu and the sine wave nonlinear regression model proposed by Chan. To gain an improved forecast by using this combined technique is usually worth considering as it is easy, and is often successful. These are very pragmatic reasons, but if the combination is successful, there are some further implications. The reason a combined forecast may be preferable is that neither constituent forecast is using all of the data in the available information set in an optimal fashion. Therefore, the success of a combination suggests that a more general model should be attempted, including the better features of the models underlying the constituent forecasts. As our proposed model gained accuracy over other univariate time series models as shown in *Table 2*, we suggest that the Singapore tourism authority applies this technique in forecasting the volume of tourist arrivals.

References

1. Chan, Y. M., Forecasting tourism: a sine wave time series regression approach. *Journal of Travel Research* 1993, **31**, 58–60.
2. Chu, F. L., Forecasting tourism arrivals: nonlinear sine wave or ARIMA? *Journal of Travel Research*, 1998 (**36**, 79–84).
3. Eadington, W. and Redman, M., Economics and tourism. *Annals of Tourism Research* 1991, **18**, 41–56.
4. Crouch, I. G., The study of international tourism demand: a survey of practice. *Journal of Travel Research* 1994, **32**, 41–55.
5. Van Doorn, J. W. M., Tourism forecasting and policy-maker—criteria of usefulness. *Tourism Management* 1984, **5**(1), 24–39.
6. Makridakis, S. and Hibon, M., Accuracy of forecasting: an empirical investigation. *Journal of the Royal Statistical Society Series A* 1979, **142**(2), 97–125.
7. Harvey, A. C., *Time Series Models*. Philip Allan, Deddington, 1981.
8. Witt, S. F. and Martin, C., Econometric models for forecasting international tourism demand. *Journal of Travel Research* 1987, **25**(3), 23–30.
9. Bates, J. M. and Granger, C. W., The combination of forecasts. *Operation Research Quarterly* 1969, **20**, 451–468.
10. Box, G. and Jenkins, G., *Time Series Analysis: Forecasting and Control*. Holden Day, San Francisco, 1976.
11. Khan, H., Phang, S. and Toh, R., Tourism growth in Singapore: an optimal target. *Annals of Tourism Research* 1996, **23**, 222–223.
12. Archer, B. H., Demand forecasting and estimation. In *Travel Tourism and Hospitality Research*, eds J. R. B. Ritchie and C. R. Goeldner. Wiley, New York, 1987, pp. 77–85.
13. Singapore Tourist Promotion Board, Annual Statistical Report on visitor arrivals to Singapore. Various issues.
14. Uysal, M. and Crompton, J., An overview of approaches used to forecast tourism demand. *Journal of Travel Research* 1985, **23**, 7–15.
15. World Tourism Organization, *Current Trend and Tourism Indicators, 3 January*. WTO, Madrid, 1994.