A COMPARISON OF THREE DIFFERENT APPROACHES TO TOURIST ARRIVAL FORECASTING

REF. NO. PQ-4556

By

Dr. Vincent Cho

Department of Management

The Hong Kong Polytechnic University

Introduction

The twentieth century witnessed a steady increase in tourism all over the world. The level of organisation, methods of transport, and the facilities available at destination points have enjoyed an accelerated pace of improvement. The arrivals of information highways and visible changes in the mass media have enhanced the appeal of tourism, and there is in greater demand of quality service at competitive prices. A smart entrepreneur in the field must know how to analyse the available data and interpret it to his or her advantage.

Forecasting plays a major role in tourism planning. The promotion of tourism projects involving substantial sums of money requires an estimate of future demand and market penetration. The commitment to developing tourism business in another geographical area would be much easier if it were possible to analyse current and past tourist traffic and predict the nature of changes in tourism demand. These extrapolative approaches to forecasting require historical data.

This paper investigates the application of three time-series forecasting techniques; namely exponential smoothing, univariate ARIMA, and Artificial Neural Networks

(ANN), to predict travel demand (i.e. the number of arrivals) from different countries to Hong Kong. Exponential smoothing and ARIMA are two commonly used statistical time series forecasting techniques. The third approach, Neural Networks, is an artificial technique in the stream of computer science. According to the analysis, Neural Networks seems to be the best forecasting method for visitor arrivals, especially those series without obvious pattern.

Background

The tourism industry, which benefits the transportation, accommodation, catering, entertainment and retailing sectors, has been blooming in the past few decades. In 1999, global international tourist receipts were around US$455 billion from an estimated 664 million tourists. According to the World Tourism Organisation (2000), international tourist arrivals worldwide are forecasted to reach 1.6 billion and tourism receipts will be over US$2 trillion by the year 2020 (1996a). The tourism industry is one of the largest growing industries. Tourism growth is particularly rapid in East Asia and the Pacific in 1999, with a growth rate of 11.5% in tourist arrivals. In 1999, tourist arrivals in the East Asia and Pacific regions reached 97.5 million, and tourism receipts were over US$73.4 billion (WTO, 2000).

Hong Kong, being part of Indo-China and Southeast Asia and the gateway to China, gains a lot from the growth of the tourism industry. In 1999 and 2000, there were 11.3 million and 13.1 million tourist arrivals respectively. In monetary terms, tourism receipts, which make the largest contribution to foreign currency revenue, were HK$56.2 billion and HK$61.5 billion in 1999 and 2000 respectively (HKTA, 2000).

As a result of the global economic downturn and the East Asian Economic Crisis in 1998, Hong Kong is experiencing decline in all industries, especially in the tourism industry. The occupancy rate of most hotels in Hong Kong averaged around 73%, while the number of tourists coming to Hong Kong from January to July 1998 dropped to 5.4 million, compared to 6.4 million in the same period in 1997 (Hong Kong Tourist Association, 1998). The largest drop, around 57.6%, was number of incoming tourists from South Korea and Indonesia (Hong Kong Tourist Association,

1998). This has had a direct effect on the retail outlets, restaurants, and hotels in Hong Kong.

In this study, tourist arrivals from six countries: the Americas, the United Kingdom, Japan, Korea, Singapore, and Taiwan are considered. With the knowledge of predicting tourist demand, we may change our promotion policy to Japanese tourists accordingly and perhaps make a bigger discount on our hotel room rates specially for for certain kind of visitors. Ultimately, the forecasting methods presented here may allow people to make more accurate forecasting of tourism and help development planning for various tourism activities.

Methodology and Analysis

This study uses exponential smoothing, seasonal ARIMA and ANN in forecasting. The following describes in detail the forecasting methods we have adopted.

Exponential Smoothing

Exponential smoothing is a forecasting method that seeks to isolate trends or seasonality from irregular variation. It has been found to be most effective when the components describing the time series may be changing slowly over time (Yaffee & McGee, 2000). In exponential smoothing, a new estimate is the combination of the estimate for the present time period plus a portion of the random error generated in the present time period. When used for forecasting, exponential smoothing uses weighted averages of the past data. The effect of recent observations is expected to decline exponentially over time.

Most visitor arrivals are affected by the economic environment and seasonal factors such as weather or public holidays. If seasonal factor exist in a series, one would expect there is large fluctuations between months. There should be a large number of visitors to Hong Kong in certain seasons like the summer vacation in the origin country. In contrast, the number of visitors may drop during the autumn season when they have fewer holidays. Thus the correlation between visitors in adjacent periods or months is expected to be low. On the other hand, most economic cycles are in terms of years, and economic indicators fluctuate slowly among months. Thus one would

expect higher correlation between adjacent quarters if economic factors strongly affect visitor arrivals. The resulting prediction curve will try to follow the actual curve with a time gap of a month. To accommodate both trend and seasonality, the Winters model uses three parameters, α, δ, and γ, to compute the forecasting. The model is given as follows,

$$\hat{Y}_{t+1}(t) = [a(t) + b(t)] \cdot Sn_{t+1}(t+1-L) + \varepsilon_t \qquad (1)$$

where $\hat{Y}_{t+1}(t)$ = the forecast for the next time period, $t + 1$;

$a(t)$ = the smoothed estimate for the level at time period t;

$b(t)$ = the smoothed estimate for the slope at time period t;

$Sn_{t+1}(t+1-L)$ = the smoothed estimate for the $(t+1)$th season made at time period $t+1-L$;

L = the periodicity of the seasonality

To update the smoothing components of the series, use

$$a(t) = \alpha \left[\frac{Y_t}{Sn_t(t-L)} \right] + (1+\alpha)[a(t-1) + b(t-1)];$$

$$b(t) = \gamma[a(t) - a(t-1)] + (1-\gamma)b(t-1); \text{ and}$$

$$Sn_{t+1}(t) = \delta \left(\frac{Y_t}{a(t)} \right) + (1-\delta)Sn_t(t-L)$$

where

α = the weighting factor for the level;

γ = the weighting factor for the trend; and

δ = the weighting factor for seasonal components.

The multiplicative Winters model consists of a linear trend and a multiplicative seasonal factor. The model assumes that the seasonal swings are proportional to the level, but the errors are not. The most recent observation is adjusted by its seasonal factor, Sn_{t+1-L}. If α is large, more weight is given to the most recent observation; this allows the level to update more quickly. For large γ, more weight is given to the most recent estimate of the slope, and small γ gives more weight to previous slope estimate. Similarly, large δ gives more weight to most recent estimate of seasonal factor and small δ would give more weight to previous estimate of seasonal factor. Each of the three smoothing parameters is updated with its own exponential smoothing equation. All together, these smoothing equations adjust and combine the component parts of the prediction equation from the values of the previous components.

In order to find the optimal values for α, γ, and δ, we simply choose the value that gives us the smallest sum of square errors (SSE) by grid search. The grid search chooses a combination for the three parameters by employing a method of trial and error. The grid values start with zero and end with one, incrementing by *0.01*. Thus the grid generates *101*101*101=1,030,301* models, with values for each parameter ranging from *0, 0.01, 0.02, ...* and so on up to *1*. The parameter that produces the smallest SSE is recorded and used to formulate models for the series.

ARIMA

ARIMA models are flexible and widely used in time-series analysis. ARIMA stands for AutoRegressive Integrated Moving Average and combines three types of processes: autoregression (AR), differencing to strip off the integration (I) of the series, and moving averages (MA). Each of the three types of process has its own characteristic way of responding to a random disturbance. The resulting general linear model (pp. 281, Pankratz 1983) is in the form:

$$\phi_p(B)\Phi_{sp}(B^L)\nabla^d\nabla_L^{sd}Z_t = \Theta_{sq}(B^L)\theta_q(B)\varepsilon_t \tag{2}$$

where Z_t *= the stationary data point;*
        B *= the backshift operator, which* $B(Z_t) = Z_{t-1}$*;*
        L *= the seasonal periodicity;*
        ε_t *= the present disturbance;*
        $\phi_p(B)$ *=* $(1-\phi_1B-\phi_2-\ldots-\phi_pB)$*, nonseasonal operator;*
        $\Phi_{sp}(B^L)$ *=* $(1-\Phi_{1L}B^L-\Phi_{2L}B^L-\ldots-\Phi_{spL}B^L)$*, seasonal operator;*
        $\theta_q(B)$ *=* $(1-\theta_1B-\theta_2B-\ldots-\theta_qB)$*, nonseasonal operator;*
        $\Theta_{sq}(B^L)$ *=* $(1-\Theta_{1L}B^L-\Theta_{2L}B^L-\ldots-\Theta_{sqL}B^L)$*, seasonal operator;*
        ∇^d *=* $(1-B)^d$*, nonseasonal differencing operator; and*
        ∇_L^{sd} *=* $(1-B^L)^{sd}$*, seasonal differencing operator.*

Identification is a critical step in building an *ARIMA(p, d, q)(sp, sd, sq)_L* model, where *p* is the AR order which indicates the number of parameters of ϕ, *d* is the number of times the data series must be differenced to induce a stationary series *Z*, *q* is the MA order which indicates the number of parameters of θ, *sp* is the seasonal AR order which indicates the number of parameters of Φ, and *sq* is the seasonal MA order

which indicates the number of parameters of Θ, and *sd* is the number of times the data series needs to be seasonally differenced to induce a seasonally stationary series.

Traditionally, identification methods are rough procedures applied to a set of data to indicate the kind of representational model that is worthy of further investigation. The specific aim here is to obtain some idea of the values of *p*, *d* and *q* needed in the general linear ARIMA model, and to obtain estimates for the parameters.

These parameters are usually determined by inspecting the behaviour of Auto-correlation Function (ACF) and Partial Auto-correlation Function (PACF) (pp. 64-67, Box et al. 1994). The ACF and PACF of a stationary series should show either a cut-off or rapidly dying pattern. In practice, the determination of *d* and *sd* requires guessing different combinations among the possible values until the desired patterns of ACF and PACF are achieved. Next will be the identification of the parameters, *p* and *q*, which involves the study of the behaviour of the ACF and PACF. On these procedures, we can establish a tentative ARIMA model. However, all parameters are determined by observation and subjective guessing, which is rather unreliable and inaccurate.

Parameter Estimation Algorithm

Upon the above drawback in estimating the parameters of the ARIMA model, we have devised an algorithm as follows:

```
For p, d, q, sp, sd, and sq each = 0 to 2
Do
      Execute SPSS ARIMA with the set parameters.
      Record the parameters and corresponding fitting error.
Until all possible combinations are tried.
Report the parameters that produce the least fitting error.
```

This algorithm tries all combinations of parameters, which are limited to an integer lying between zero and two. The combination with the least fitting error will be searched. The range limitations of the parameters are set to restrict the search to a reasonable scope. Parameters greater than two are rarely used in practice.

For example, the series of visitors from the USA was modelled as AR order $p=2$, MA order $q=0$, $sp=2$ and $sq=1$ with differencing $d=0$ and seasonal differencing $sd=1$. The corresponding fitting error is the lowest among all different combinations of parameters. Moreover, we also restricted the solution space so that the estimated coefficients are all within a predetermined confidence limit of 95%.

Artificial Neural Network (ANN)

Elman's ANN model (Elman, 1990) is used to predict arrivals. Elman's Network is a recurrent network, which the output of hidden layers is feedback to itself, thus it is especially suitable for fitting time series (Jhee & Lee, 1996). The activations in the hidden layer at time $t-1$ are copied into the context vector, which is the input to the network for time t. This is equivalent to having the hidden layer completely and recurrently connected, and back-propagating one step in time along the recurrent connections. Therefore, the reaction of the network to the new input is a function of both the new input and the preceding context. What is stored in the context vector at any given time is a compressed trace of all preceding inputs, and this compressed trace influences the manner in which the network reacts to each succeeding input.
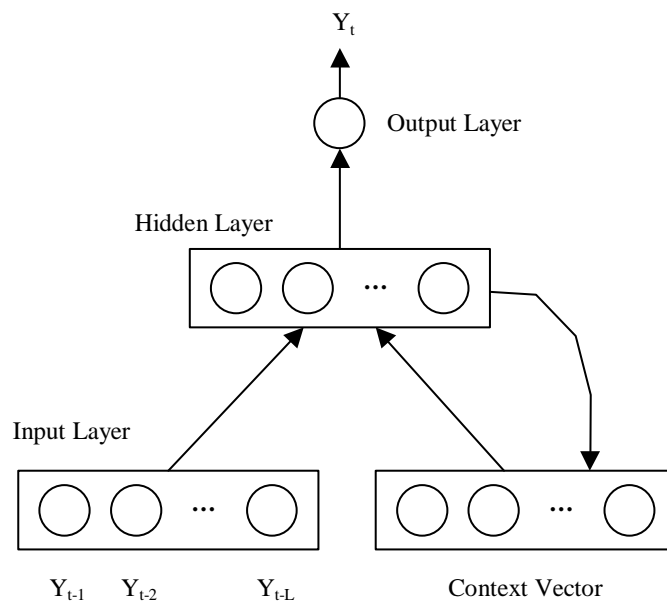

Figure 1: Elman Network

Data Collection

Tourist arrival statistics in Hong Kong were obtained from the yearly reports of the HKTA, and countries such as the USA, the UK, Japan, Korea, Singapore and Taiwan (as shown in Figure 2) were considered because they contribute the most visitors from all over the world to Hong Kong. The period of collected data is from January of 1974 to December of 2000.
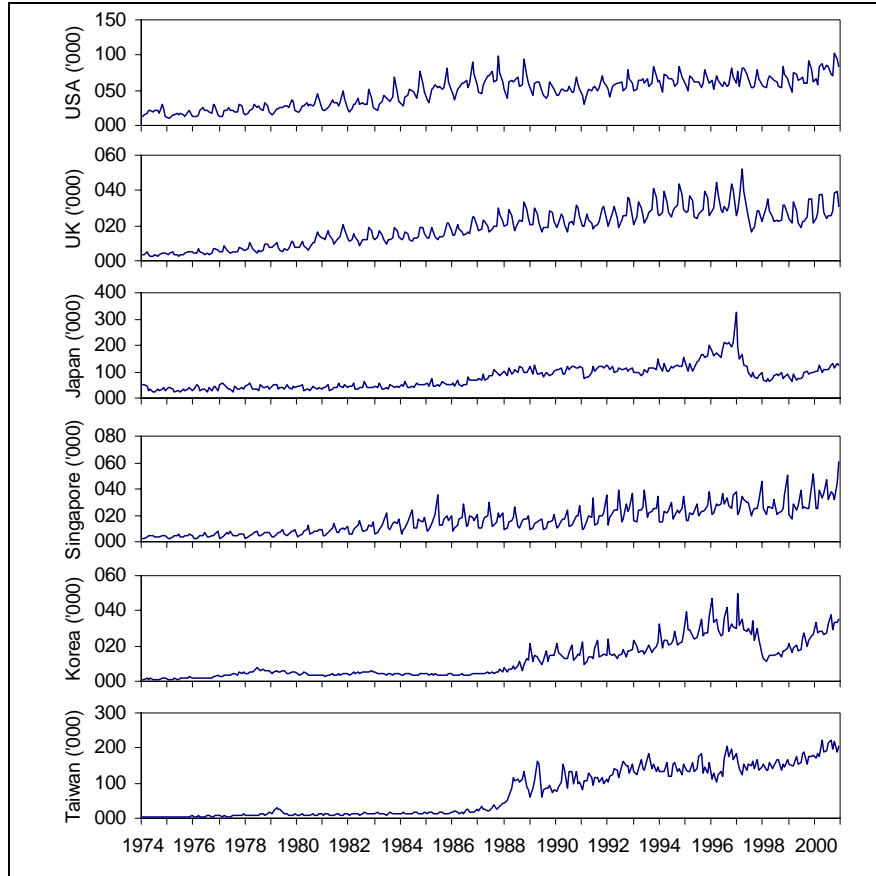

Figure 2: Visitor Arrivals

Monthly statistics of tourist arrivals from the mainland China has been recorded only after the nineties. Though more and more tourists from mainland China travel to Hong Kong, China is not included in this paper as it does not have adequate data to allow us to build models in the analysis.

The collected data were divided into two sets, training data and testing data, in order to testify the performance of the three suggested forecasting methods. To achieve a more reliable and accurate result, a long period served as the training period. Based on these considerations, we selected the period from January of 1974 to December of

1998 as the training period, and January of 1999 to December of 2000 as the testing period.

Results and Discussion

Exponential smoothing

The parameters of α, γ, and δ got by grid search are shown in Table 1 for the exponential smoothing. The parameter α smoothes the level equation. While the value is low, it indicates that the corresponding series is quite steady over the training period; on the other hand, high value indicates the fluctuation of arrivals is large. As shown in Table 1, the parameters of trend smoothing, γ, for all series are zero. It indicates the slopes of the six series are constant over the training period, or the trend can be represented by a straight line. Seasonal effects are captured by the seasonal parameter, δ, in the seasonal equation. These parameters α, γ, and δ are all interrelated. For example, large value of δ will tend to have low value of α and vice versa.

From Table 1, the pattern of UK and Taiwan are similar, The α values are high and the δ values are zero. One would thus suspect that these countries' tourist arrivals depend strongly on the most recent figure of arrivals, with constant trend and steady seasonal effect. On the other hand, the exponential smoothing model with low α and high δ values indicate that the series of USA, Japan, Korea, and Singapore may be strongly affected by seasonal factors

Table 1: Parameters of Exponential Smoothing

	Alpha (General)	Gamma (Trend)	Delta (Seasonal)
USA	0.36	0.00	0.74
UK	0.80	0.00	0.00
Japan	0.52	0.00	0.99
Korea	0.60	0.00	0.97
Taiwan	0.82	0.00	0.00
Singapore	0.22	0.00	0.68

ARIMA

Using the training period, the estimation of parameters using the described searching algorithm is shown in Table 2. With these parameters, the coefficients in equation *(2)* were deduced (Table 2).

One common factor $\phi_1 Z_{t-1}$ is all non-zero among all countries. This would support that the previous month has an effect on the current one. Except Japan, all series need either differencing or seasonal differencing to produce stationary series for analysis. The series of Korea requires differencing; while the series of USA, UK, Taiwan and Singapore requires seasonal differencing. Since the Japan series needs no differencing, its model looks the most complicated (p=2, q=2, sp=1, and sq=2).

Table 2: ARIMA Models

	USA	Japan	Taiwan	Korea	UK	Singapore
d	0	0	0	1	0	0
sd	1	0	1	0	1	1
Constant μ	2018.50*	70806.44*	6782.81**	71.01*	1088.50**	1069.10**
ϕ_1	0.63**	1.93**	0.84**	0.94**	0.78**	0.83**
ϕ_2	0.21**	-0.94**				
θ_1		1.23**		1.39**		0.49**
θ_2		-0.29**		-.39**		
Φ_{1L}	-0.98**	0.92**	0.31**	0.63**		
Φ_{2L}	-0.47**					
Θ_{1L}	-0.51**	0.73**	0.73**		0.47**	0.29**
Θ_{2L}		-0.21*				0.15*

Artificial Neural Networks

All the ANN models for the six series are with the same structure as stated above. Inputs of the series are normalised to the range (-1, 1) before feeding into the networks. The training processes of the recurrent networks stopped when the sum of square errors (SSE) reach 0.005 in 500 epochs (cycles). The error rate 0.005 is chosen to achieve a good accuracy and avoids overfitting the model.

Comparisons among the three forecasting methods

The predictions of visitor arrivals in the testing period are done using the three forecasting methods, and Table 3 illustrates our results. To compare the prediction

performance of the three approaches for the period January 1999 to December 2000, the following measures of accuracy were calculated: root mean square error (RMSE) and mean absolute percentage error (MAPE) (Gaynor and Kirkpatrick, 1994, pp.13-14).

RMSE is defined as   MAPE is defined as

$$RMSE = \sqrt{\frac{\sum e_t^2}{n}}$$

$$MAPE = \frac{\sum {|e_t|}/{Y_t}}{n}$$

where e_t is the forecast error in the time period t; Y_t is the actual value in time period t; and n is the number of forecast observations in the estimation period.

Table 3: Performance of three forecasting methods

	ARIMA		Exp Smoothing		Neural Network	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
USA	9434.06	10.23	9494.08	10.29	**8678.94**	**10.11**
Japan	45487.13	44.52	32846.00	26.43	**11154.86**	**10.32**
Taiwan	21267.28	8.24	29911.08	13.02	**20089.23**	**8.00**
Korea	10109.81	32.13	7330.44	21.18	**2735.55**	**9.32**
UK	5124.45	15.73	**2699.98**	**7.90**	4385.30	13.32
Singapore	5944.24	14.26	6206.62	14.20	**5064.26**	**11.99**

* the minimal errors are highlighted in bold

The results of ANN, with more hits of minimal errors, seem to give the best performance of the three methods. In particular, this is the best predictor for countries like the Japan and Korea. These two series show no obvious pattern in Figure 2, where usual time series methods can do a little on such kind of data. ANN outperforms in forecasting all series except for that of UK. For visitor arrivals from UK, Exponential Smoothing seems to be the best predictor. This may not be coincidence, as shown in Figure 2, visitor arrivals from UK, Japan and Korea dropped heavily at around the years 1997 and 1998, which is the time of economic crisis in Asia; but increased again after the crisis. The series repeated their patterns and scales which had occurred in several years ago. While exponential smoothing models are basically weight averages of past observations, it is not surprising that the method performs well for these three countries. Indeed, the accuracy of forecasting these three series by exponential smoothing is better than forecasting by ARIMA.

Among the three forecasting methods, ANN, although its training time is the slowest, is the highest accurate one. This indicates that ANN is more suitable and can be applied to the prediction of fluctuating series of visitor arrivals.
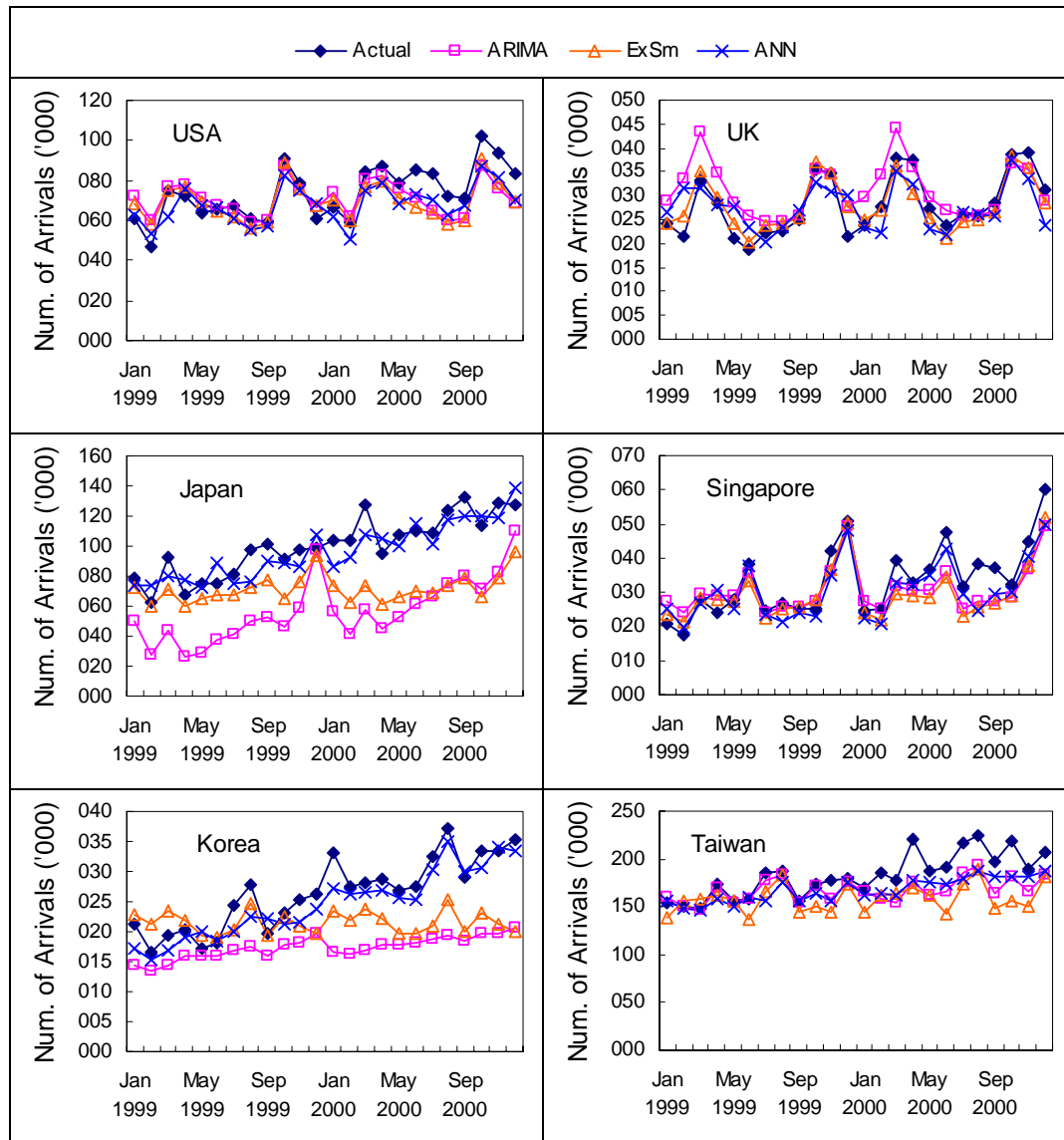

Figure 3: Forecasts of Visitor Arrivals

Conclusions

This paper investigates three forecasting methods: exponential smoothing, ARIMA and Artificial Neural Networks (ANN) and applies them to predict visitor arrivals from six countries (the USA, the UK, Singapore, Japan, Taiwan, and Korea) to Hong Kong. It was found that ANN outperforms exponential smoothing and ARIMA. ANN performs the best forecasting for all visitor arrival series except UK, which the best

method for it is exponential smoothing. Thus we can conclude that ANN is good for predicting visitor arrivals, especially for those series with less obvious patterns like Japan and Korea; otherwise, ARIMA and exponential smoothing are sufficiently adequate.

With an accurate estimate, the Hong Kong government and the hospitality industry can perform better tourism strategic planning. In business, there should be greater freedom and stability to allow more corporations to come to Hong Kong to establish their representative offices.

In future, more research will be focused to apply more sophisticated forecasting techniques using latest technology such as Genetic Algorithm (Duda et al., 2001), Fuzzy Neural Systems (Skapura, 1995), Probabilistic Rules (Wüthrich, 1996), or some other advanced data mining techniques to predict visitor arrivals. Concerning Hong Kong, special attention should be put on the area of visitor arrivals from China as it is emergingly important after the change over in 1997.

Acknowledgement

References

Aleksander and Morton (1990). *Introduction to Neural computing*, North Oxford Press.

Andrew, W.P., Cranage, D.A., and Lee, C.K. (1991). Forecasting Hotel Occupancy Rates with Time Series Models: An Empirical Analysis, *Hospitality Research Journal*, pp. 173-181.

Beale, R. and Jackson, T. (1990). *Neural Computing – An Introduction*. Bristol: Adam Hilger.

Box, G.E.P., Jenkins, G.M., and Reinsel, G.C. (1994). *Time Series Analysis, Forecasting and Control*, Prentice Hall.

Briassoulis, H. (1991). Methodological issues: Tourism Input-Output Analysis, *Annals of Tourism Research*, Vol. 18, pp.485-494.

Duda, R.O., Hart, P.E., and Stork, D.G. (2001), *Pattern Classification*, 2nd Edn, New York: John Wiley & Sons, Inc.

Elamn, J.L. (1990). Finding Structure in Time, *Cognitive Science*, 14, 179-211.

Eby, F. and O'Neill, W. (1977). *The Management of Sales Forecasting*, Lexington, MA:D.C. Heath and Company, p. 254.

Farnum, N.R. and Stanton, L.W. (1989). *Quantitative Forecasting Methods*, PWS-Kent Publishing Company.

Fletcher, J.E. (1989). Input-Output Analysis and Tourism Impact Studies, *Annals of Tourism Research*, vol. 16. Pp.514-529.

Gaynor, P.E. and Kirkpatrick, R.C. (1994). *Introduction To Time-Series Modeling And Forecasting In Business And Economics*. NY: McGraw-Hill, Inc.

Haugh, I.D. (1976). Checking the Independence of Two Covariance-Stationary Time Series: A Univariate Residual Cross-Correlation Appraoch, *Journal of the American Statistical Association*, vol. 71, pp. 378-485.

Hodgetts, R.M. and Luthans, F. (1997). *International Management*, McGraw Hill, Third edition, pp. 211-236.

Hong Kong Tourist Association. *Tourism Receipts* – various issues 1986-2000.

Hong Kong Tourist Association. *Visitor Arrival Statistics* – various issues 1974-2000.

Jhee, W.C., and Lee, J.K. (1996), "Performance of Neural Networks in Managerial Forecasting" in Trippi, R.R. and Turban, E., *Nerual Networks in Finance and Investing*, Revisited Edn., pp.703-733, Chicago: IRWIN.

Layton, A.P., Defris, L. V., and Zehnwirth, B. (1986). An International Comparison of Economic Leading Indicators of Telecommunication Traffic, *International Journal of Forecasting*, Vol. 2, pp. 413-425.

Pankratz, A. (1983). *Forecasting With Univariate Box-Jenkins Models*. NY: John Wiley & Sons.

Sheldon, P.J. and Var, T. (1985). Tourism Forecasting: A Review of Empirical Research, *Journal of Foreasting*, Vol. 4, pp. 183-195.

Skapura, D.M. (1995). *Building Neural Networks*, New York: ACM Press.

Smeral, E. and Witt, S.F. (1996). Econometric Forecast of Tourism Demand to 2005, *Annals of Tourism Research*, Vol. 23, pp. 891-907.

Smeral, E., Witt, S.F., and Witt, C.A. (1992). Econometric Forecasts: Tourism Trends to 2000, *Annals of Tourism Research*, Vol. 19, pp.450-466.

Stone, J.R.N. (1954). Linear Expenditure Systems and Demand Analysis: An Application to the pattern of British Demand, *Economic Journal*, Vol. 64, pp.511-527.

Turner, L, Kulendran, N., and Fernando, H. (1997). the Use of Composite National Indicators for Tourism Forecasting, *Tourism Economics*, Vol. 3, No. 4, pp.309-317.

Wagner, J.E. (1997). Estimating the Economic Impacts of Tourism, *Annals of Tourism Research*, Vol. 24, No.3, pp.592-608.

Wander, S.A. and Erden, J.D. (1979). Estimating the Demand for International Tourism using Time Series Analysis, *Proceedings of the International Symposium on Tourism in the Next Decade*. Washington, DC.

Wei, W.W.S. (1990). *Time Series Analysis - Univariate and Multivariate Methods*. NY: Addison-Wesley Publishing Company, Inc.

Weller, B.R. (1990). Predicting Small Region Sectoral Responses to Changes in Aggregate Economic Activity: A Time Series Approach, *Journal of Forecasting*, Vol. 9, pp.273-281.

Wüthrich B. (1996). Probabilistic Knowledge Bases, *IEEE Transactions of Knowledge and Data Engineering*, Vol. 7, No. 5, pp. 691-698.

World Tourism Organization. (1996a). *Tourism 2020 Vision*, Madrid, Spain

World Tourism Organization. (1996b). *Tourism Highlights 1996*, Madrid, Spain.

World Tourism Organization. (2000). *Tourism Market Trends: East Asia and the Pacific, 2000 Edition*, Madrid, Spain.

Yaffee, R.A., and McGee M. (2000). *Introduction to Time Series Analysis and Forecasting*, San Diego: Academic Press, Inc.

Zhou, D., Yanagida, J.F., Chakravorty, U., and Leung, P.S. (1997). Estimating Economic Impacts from Tourism, *Annals of Tourism Research*, Vol. 24, No.1, pp. 76-89.