Golden Week Tourist Flow Forecasting Based on Neural Network

Kailiang Wu[1], *student member, IEEE,* Bin Dai[2]

Abstract- **Golden Week is a collection of national holidays within seven days. Accurate forecast of tourist flow will boost the business of tourism and optimize the allocation of resources. In this paper, taking Chinese Golden Week as a case study, we implement a forecasting procedure based on a hybrid model using Kalman Filter and Neural Network. The result of this technique is evaluated and compared with other common forecasting models. We conclude that the hybrid model is effective and outperforms the other methods.**

Keyword: Golden Week, Tourist Flow Forecasting, Kalman Filter, Neural Network

I. Introduction

The Golden week, which combines national holidays with well-placed weekend, is a well known profitable tourism market in many countries like China and Japan. As a country with the largest population in the world, China now has the biggest tourism market especially in the Golden Week from May 1st to 7th and from Oct1st to 7th because of the nation-wide holidays for Chinese people together with foreign tourists. The huge market has been boosting the business of national tourism. However, the large flow of the tourists will be a heavy burden for airports, hotels and sightseeing spots during the golden week. Therefore, an optimization of the allocation of social resources in such areas is crucial and correct forecasting is a prerequisite of the optimizing process which is the main objective of this paper.

There are large amount of researches done in the forecasting of traffic flow and electric road [1], [2]. The main methods that are including in these researches are: traditional regression analysis; time series model such as Autoregressive Moving Average (ARMA) and Kalman Filter [3], [4]; Gray Model; Exponential Smoothing [14]; Artificial Neural Network (ANN) Model. Due to the nonlinear nature of flow prediction, technique of ANN has been widely paid attention to. ANN is also applied in plenty of areas [5] including forecasting, pattern recognition and classification.

Previous researches used ANN or regression model to forecast the flow of railroad or airline passengers, other applications of ANN in prediction are also proved to be effective. However, few researches have been implemented in the forecasting of the tourist flow, especially in Golden Week. The objective of this paper is to apply current well-known forecasting models of ANN and Kalman Filter to forecast the tourist flow of during Chinese Golden Week. In section 3 and 4, the forecasting models using Kalman Filter and ANN is proposed respectively. We come up with a hybrid model using both of the methods in section 5 and followed by comparison of models.

II. Problem Formulation

A. *Analysis of Tourist Flow*

The data is the fundamental element of our research since they serve as the information set for design and training. Our data, which is offered by China National Tourism Administration [6] includes the daily tourist flow in most of *Chinese Famous Sightseeing Spots*（CFSS）(e.g. The Great Wall, The Imperial Palace) on the Golden Week of National and May Holiday from 2000 to 2006. The data in 2003 is neglected because of the invasion of SARS which diminishes the relevance of the traveling behavior.

The data presents the following characteristics:

- Yearly periodicity. The tourist flow shows periodically pattern, which is primarily due to the similar traveling behavior of people year after year. The maximum flow (the peak of the curve) occurred in the middle of the week.
- Fluctuation of daily flow in different years. External factors, such as weather condition, congestion and government policy, exert significant influence on the variation of tourist flow. Such fluctuation is regarded as the environmental noise of the model.

We take the Correlation analysis [7] on the data set to study how the previous data affect the present tourist flow. In **Fig1** and **Fig2** (The graph is processed with a continuous scale by EXCEL), It's obvious that the present tourist flow is highly correlated with that of previous day. In addition, the data of the same day in the previous year is also an influential factor.

B. *Benchmark*

The formulas used in the error analysis are the Average Absolute Error (AAE) and Maximum Error (ME). AAE calculates the average absolute forecast error value of a certain period. It follows:

[1] Kailiang Wu is with the Department of Information Science and Engineering, Zhejiang University, 310027,China. Email: kailiangwu@gmail.com

[2] Bin Dai is with the Department of Science, Zhejiang University, 310027, China. Email: daibin5@gmail.com

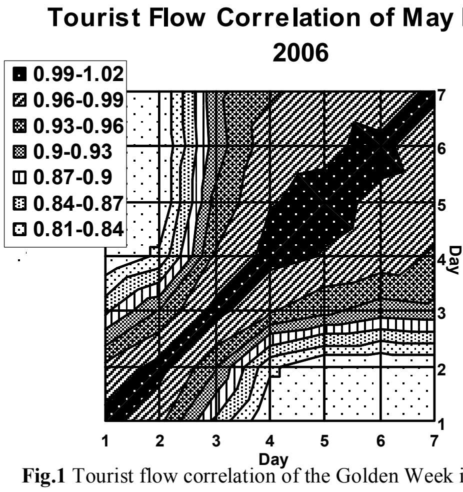

Fig.1 Tourist flow correlation of the Golden Week in 2006

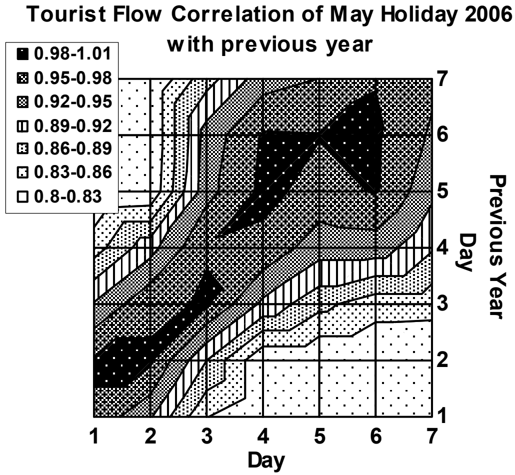

Fig.2 Tourist flow correlation with previous year

$$e_{avg} = \frac{\left|\sum_{i=1}^{7} \hat{y}_i - y_i\right|}{7} \qquad (1)$$

Where $\hat{y}$ is the forecast value and y is the real value. The number 7 in the denominator means the number of day during the "Golden Week". The ME is the largest absolute error occurs in the forecast of the same period of time. It is:

$$e_{\max} = Max[|\hat{y}_1 - y_1|,|\hat{y}_2 - y_2|...,|\hat{y}_7 - y_7|] \qquad (2)$$

This criterion will be used in evaluating the effectiveness of our forecast model presented in the following sections.

III. Forecasting Model Based on Kalman Filter

Based on the characteristic of data, the tourist flow expected at a particular day (state) can be predicted by filtering out fluctuation from flow value observed at corresponding day in the previous year. The Kalman Filter provides the estimation of the process state through a recursive algorithm, which minimizes the mean of squared error. Firstly, the state space-form model, which is obtained through system identification technique, is given as follow:

$$x(n+1) = Ax(n) + w(k) \qquad (3)$$

$$y(n) = Cx(n) + v(n) \qquad (4)$$

Where:
x: the state vector
y: observed output
A: state transition matrix
C: observation vector
w: vector of white noise with Covariance Q
v: white random scalar with variance R

$$Q = E[w(n) \times w(n)^T] \quad R = E[v(n) \times v(n)^T]$$

Then, the Kalman Filter obtains the optimal state estimation by calculating the recursive update solution:

$$K(n) = \frac{P(n)C^T}{CP(n)C^T + R} \qquad (5)$$

$$\hat{x}(n+1) = A\hat{x}(n) + K(n)[y(n) - C\hat{x}(n)] \qquad (6)$$

$$P(n+1) = A[P(n) - K(n)CP(n)]A^T + R \qquad (7)$$

Where:
$K(n)$: Kalman Gain at time n
$P(n)$: Prediction Covariance
$\hat{x}(n+1)$: Optimal estimate of state n+1 given the information up to n

The optimal output prediction will be:

$$\hat{y}(n+1) = C\hat{x}(n+1) \qquad (8)$$

Shanghai Oriental Peal (SOP) is a famous attraction and the symbolic architecture of the city. We take the tourist flow data in SOP during Golden Week as the test of our first model.

The forecast result of SOP is showed in **Fig.3**. The unit of the data in the y-axis is ten thousand (10^4) people. The Kalman Filter based model works well in tracking the general trend of the tourist flow. However, the error occurred is high shown by table 1.

TABLE I
THE FORECAST ERROR OF KALMAN BASED MODEL

Year	Average Absolute Error	Maximum Error
2004	0.3694	0.5787
2005	0.3953	0.6803
2006	0.4872	0.8138

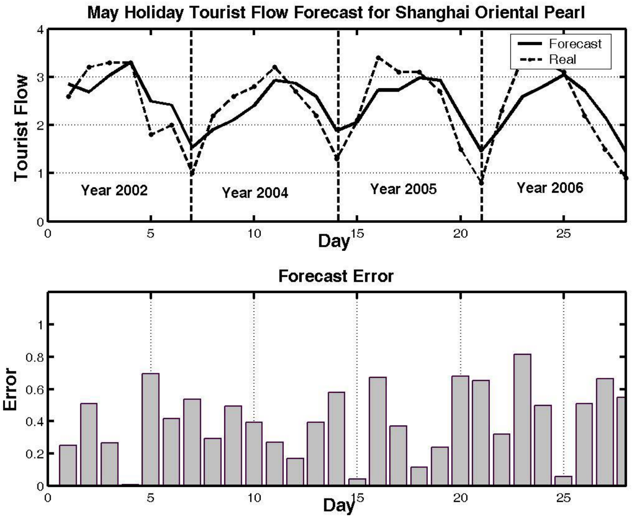

Fig.3 Tourist Flow Forecast for Shanghai Oriental Peal Based on Kalman Filter

IV. FORECASTING MODEL BASED ON ANN

A General Design of the network

In this study, multi-layer perceptron (MLP) neural network models are used based on a class of supervised learning algorithm termed back-propagation (BP). The BP neural network has been a popular model that can approximate the time-series data. There are four input units and one output unit as showed in **Fig.4**. The number of units in the hidden layer is 15. We found that the model with linear transfer function performs better than the commonly used Sigmoid function.

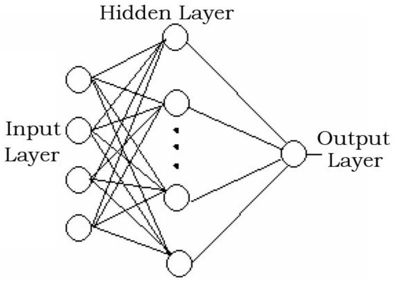

Fig.4 Topologic architecture of ANN

B. Input Design

There are many factors that may influence on the result of the forecast of tourist flow, such as historical data, the weather Information, etc. We use the following factors as the component of input unit.

1) *Historical data*

The tourist will carefully plan his/her trip before the holiday. They certainly do not like crowed place and probably incline to visit new popular attraction. So the historical popularity of the CFSS exerts profound influence on the flow of the predicting day. In terms of the correlation analysis in section 2, the tourist flow of previous day and last year is taken into account as an important model input.

2) *Traveling intention of the tourists*

According to a nation-wide survey [9] about the traveling plan of the visitors in Chinese Golden Week, 33.3% of the citizens choose to travel in the first three days, among which, 83.3% choose to visit from May 2nd to 4th; in addition, about 28% of the citizens would love to have 5 days trip in the Golden Week. In other word, in the first few day of the holiday, most of the tourists are not intended to have a trip. However, in the following days, the tourists are active so their intention undergoes an increase. As the end of the holiday approaches, their intention would decrease because they have to prepare to work or go back to their home cities. In this study, we take into consideration the intentions to visit of the tourists and use a quadric curve to fit this component. The value of it is a real number in [0,1], with a maximum in the third and fourth day in the golden week.

3) *Weather Information*

The weather of CFSS obviously has great impact on the tourist flow. Based on data of weather report, we evaluate the weather in five ranks denoted -1, -0.5, 0, 0.5, 1 as one input unit representing that the weather in the predicting day is extremely bad, bad, no effect, quite good and great.

4) *Congestion Factor*

From the data of the tourist flow we find that most popular places such as the Imperial Palace are overcrowding which makes the tourists bored with the congestion. Thus we assess of the congestion in the previous day and in the last year as a factor influencing the flow of the present day. The congestion coefficient denoted as C is defined by the percentage of tourist flow over the maximum capacity of CFSS, which is probably greater than 1, for example in Imperial Palace C is about 300%. The input unit of the network is obtained by the congestion coefficient as follow:

If C≥1, then the input would be $\frac{1}{C}-1$;

Otherwise, the input would be 1-C;

In this definition, the less crowded of CFSS, the more eager visitors are to have a trip; on the other hand, great congestion prevents the tourists from visiting the CFSS.

C. Network Training Procedure (BP)

The training set includes four years (2000-2003) data and the rest are taken to be the test set. Initially, the weights in the network were assigned randomly. The performance was gradually improved by changing the weights using the back-propagation learning algorithm.

TABLE II
THE FORECAST ERROR OF ANN MODEL

Year	Average Absolute Error	Maximum Error
2004	0.3127	0.6433
2005	0.2510	0.5554
2006	0.1832	0.3408

D. Results

Fig.5 demonstrates the tourist flow forecast of SOP from 2004 to 2006. The Neural Network based model improves the forecast precision. But it is noted that the model appears to be incapable of grasping the general trend. Specifically, the forecast precision is not satisfactory enough when it comes to the peaks (the middle of the week) of the flow curve. This is because people show irregular traveling patterns due to the transportation, tourism service and facility problems caused by great congestion during the peak days [15].

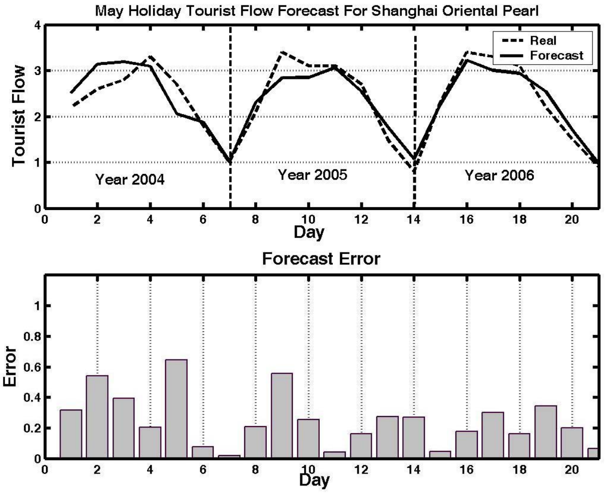

Fig.5 Tourist Flow Forecast for Shanghai Oriental Peal By ANN

E Analysis

To assess the weight of the four input components on the overall forecast precision, we perform the ANN model four times by omitting each of the input to see the change of forecast precision.

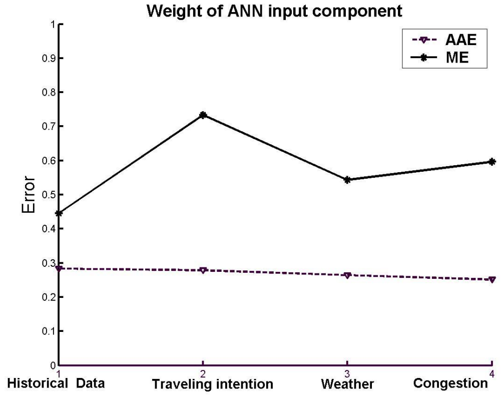

Fig.6 Analysis of ANN input component

Besides the historical data, the traveling intention factor has the most profound impact on the overall forecast precision of ANN model. On the other hand, the tourist flow time series is complex and contains strong nonlinearity due to the irregularity and uncertainty of people's intension. Therefore, the nonlinear nature of traveling intention directly yields poor forecast performance of ANN during the middle of the golden week.

The higher of the tourist flow, the greater of congestion and thus leads to irregular traveling pattern which finally caused low precision of ANN model during peak period. Hence, we attempt to figure out the nonlinearity via eliminating the uncertainty in the next section.

V. Hybrid Model Based on Kalman Filter and ANN

To take full advantage of individual strength of Kalman Filter and ANN, we develop a hybrid model integrating both of the techniques.

A. The design of The Hybrid Model

In the first model, Kalman Filter proves to be good at generating the smooth trend of tourist flow. Thus we use the model to forecast the flow variation value [10] which is taken as a new input component of the ANN.

$$\Delta y(n) = y(n) - y(n-1) \qquad (9)$$

The prediction output of Kalman Filter changes to:

$$\Delta \hat{y}(n+1) = C\{A\hat{x}'(n) + K(n)[\Delta \hat{y}(n) - C\hat{x}'(n)]\} \qquad (10)$$

This component serves as a *trend predictor.*

In ANN model, the function of the flow that is to be forecasted is written as:

$$\hat{y}(t) = f[\omega_1(t), u(t)] + \varepsilon(t) \qquad (11)$$

In which the network is trained and adapted using the set of inputs and outputs $\{u(t), y_{real}(t)$: t=1,2 … T}.

Based on the analysis previously, we observe that $\varepsilon(t)$ is high during the peak period due to the uncertain traveling pattern. In this sense, $\varepsilon(t)$ fluctuates with regard to the ANN forecasting value $\hat{y}(t)$. The larger of the tourist flow, the higher of error. Therefore we denote a new function of error by:

$$\varepsilon(t) = g[y(t)] \qquad (12)$$

Since the typical properties of this function are unknown, we utilize the capability of neural network in fitting nonlinear function. Here, a second ANN$_2$ (*error estimator*) connecting in series with the output of the first is used, which is described as:

$$\hat{\varepsilon}(t) = g[w_2(t), \hat{y}(t)] + e(t) \qquad (13)$$

Where $\hat{\varepsilon}(t)$ is the estimation of error occurred in the ANN$_1$ with respect to the output $\hat{y}(t)$. The two ANN eventually form the neural network system and the ultimate forecast of tourist flow is updated as:

$$\hat{Y}(t) = \hat{y}(t) - \hat{\varepsilon}(t) \qquad (14)$$

This hybrid model is designed to create synergetic effect on eliminating the uncertainty and further improving the forecast precision.

B. The Topologic Architecture of the Hybrid Model

The neural network system in this model is shown in the **Fig.7**. The ANN$_1$ is similar with the former introduced ANN model but being added a new input unit called Kalman *trend* *predictor*. The second part ANN$_2$ is designed to update the forecast by the mechanism of *error estimator*.

The original training set is divided into two parts in which some are taken to be the training set of ANN$_1$ and the other to be the test set of it. The forecast error of this test set is put to be the expected outputs of ANN$_2$ in which the forecast value of ANN$_1$ is the input. The rest of data set is used to testify the performance of the whole system.

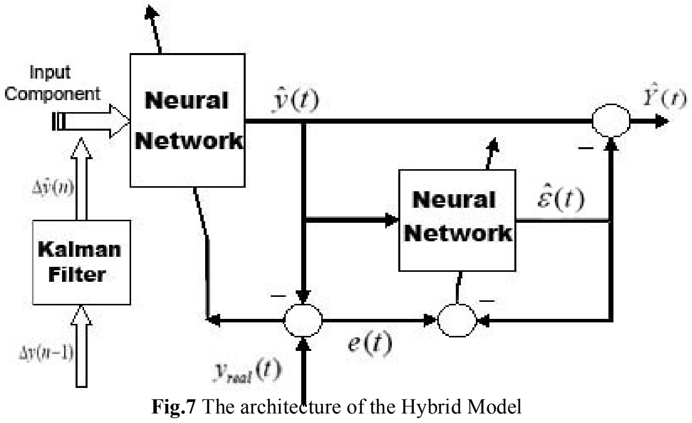

Fig.7 The architecture of the Hybrid Model

TABLE III

THE FORECAST ERROR OF HYBRID MODEL

Year	Average Absolute Error	Maximum Error
2004	0.1067	0.2666
2005	0.1847	0.4477
2006	0.1715	0.4439

C. *Results*

In **Fig.8**, the overall forecast can be an excellent proximity of the real data. The hybrid model technique can successfully capture the nonlinearity pattern in the tourist flow. Appling this technique to other places, the forecast remains precise, reflecting the wide application of our model. The result of West Lake, the heaven-like scenery in Hangzhou, is shown in the **Fig.9**.

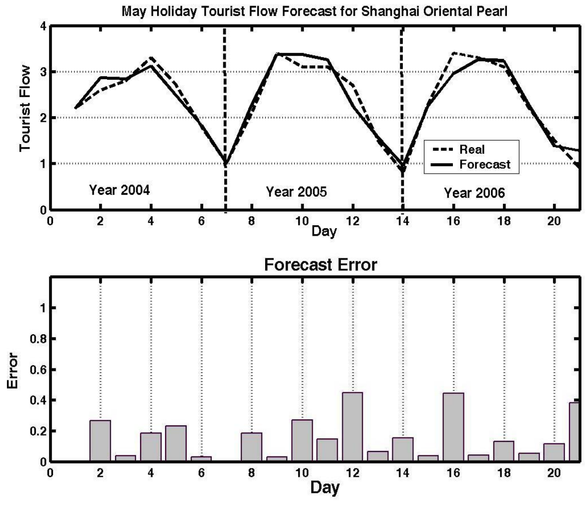

Fig.8 Tourist Flow Forecast for Shanghai Oriental Peal By Hybrid Model

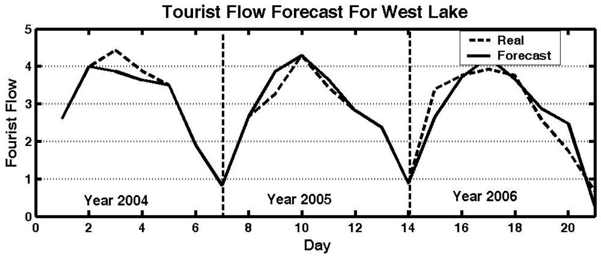

Fig.9 Tourist Flow Forecast for West Lake

VI. COMPARISON OF MODELS

The commonly used regression and ARMA models are first presented here for the comparison. To validate the effectiveness of hybrid model, we also compare its result with the average of two independently trained ANN.

The Average Absolute Error of the 5 forecasting models are given in **Fig.10.** It is observed that the AAE (the black bar) for hybrid model is considerably lower than common models and pure Kalman and ANN techniques, which proves the best performance of hybrid model.

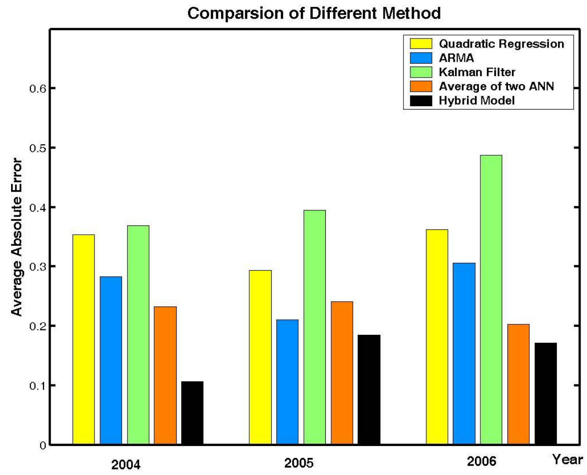

Fig.10 Comparison plot of AAE

VII. CONCLUSION

The forecast of tourist flow in CFSS during golden week is yearly periodical with fluctuation. In this paper, we propose a hybrid model incorporating the Kalman Filter and MLP Neural Network. Their synergetic effect shows promising performance. This model is also flexible when applying extensively in different CFCC.

Future work is expected for further validation of our hybrid model when the future data is available. To provide guidance for other countries with similar one week holidays, a more generalized case study will be researched.

REFERENCES

[1] D.W. Bunn and E.D Farmer, "Comparative model for eletrical load forecasting," *John Wiley & Sons Ltd.* 1985

[2] Chao Han, Su Song, "A review of some main models for traffic flow forecasting" *Intelligent Transportation Systems, 2003*. Proceedings. 2003 IEEE Volume: 1 12-15 Oct. 2003, Page(s): 216- 219 vol.1

[3] Moghram, S.Rahman, "Analysis and evaluation of five short-term load forecasting techniques", *IEEE Trans. on Power Systems*, Vol.4, No.4, October 1989

[4] G.Gross and F.D.Galiana, "Short term load forecasting," *Proc. IEEE*, Vol.75, No.12, Page:1158-1573

[5] Y.H.Pao, D.J. Sobajic, "Current status of artificial neural network applications to power systems in the United States," *T.IEE Japan*, vol111-B, n.7, '91,Page: 690-696

[6] China National Tourism Administration, http://www.cnta.com/lyen/index.asp

[7] Jianxin Xu, Lim Wei Ping, "Net flow modeling and forecasting for Honda Dirrac", *in press*

[8] Greg Welch,Gary Bishop,An Introduction to the Kalman Filter

[9] Chongqing News, http://www.cq.chinanews.com/newsview.asp?nid=78231,April, 2006

[10] Gastaldi, M.; Lamedica, R; "Short-term forecasting of municipal load through a Kalman filtering based approach" *Power Systems Conference and Exposition, 2004. IEEE*, Page(s): 1453- 1458 vol.3

[11] Yaming Ma, Luh, P.B, " A neural network-based method for forecasting zonal locational marginal prices" *Power Engineering Society General Meeting, 2004. IEEE* 6-10 June 2004 Page(s): 296- 302 Vol.1

[12] Trudnowski, D.J, cReynolds, W.L.; Johnson, J.M."Real-time very short-term load prediction for power-system automatic generation control" *IEEE Transactions on Control Systems Technology*.Volume: 9 Issue: 2 Mar 2001, Page(s): 254-260

[13] Feng Gao; Xiaohong Guan; Xi-Ren Cao; Papalexopoulos, A. "Forecasting power market clearing price and quantity using a neural network method" *Power Engineering Society Summer Meeting, 2000. IEEE* Volume: 4 2000 Page(s): 2183-2188 vol. 4

[14] Gardner, E.S.: Exponential Smoothing: The State of the Art. Journal of Forecasting 4(1985) 1-28

[15] Tang Hanyin, “Problem exit in China tourism golden week and its countermeasure research”, *Special Zone Economy, Issue:1, 2006, Page 206-207*