Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 31 (2014) 842 – 851

Procedia Computer Science

2nd International Conference on Information Technology and Quantitative Management, ITQM 2014

A Novel Divide-and-Conquer Model for CPI Prediction Using ARIMA, Gray Model and BPNN

Yudie Du, Yue Cai, Mingxin Chen, Wei Xu*, Hui Yuan, Tao Li

School of Information, Renmin University, Beijing, 100872, China

Abstract

This paper proposes a novel divide-and-conquer model for CPI prediction with the existing compilation method of the Consumer Price Index (CPI) in China. Historical national CPI time series is preliminary divided into eight sub-indexes including food, articles for smoking and drinking, clothing, household facilities, articles and maintenance services, health care and personal articles, transportation and communication, recreation, education and culture articles and services, and residence. Three models including back propagation neural network (BPNN) model, grey forecasting model (GM (1, 1)) and autoregressive integrated moving average (ARIMA) model are established to predict each sub-index, respectively. Then the best predicting result among the three models' for each sub-index is identified. To further improve the performance, special modification in predicting method is done to sub-CPIs whose forecasting results are not satisfying enough. After improvement and error adjustment, we get the advanced predicting results of the sub-CPIs. Eventually, the best predicting results of each sub-index are integrated to form the forecasting results of the national CPI. Empirical analysis demonstrates that the accuracy and stability of the introduced method in this paper is better than many commonly adopted forecasting methods, which indicates the proposed method is an effective and alternative one for national CPI prediction in China.

Selection and peer-review under responsibility of the Organizing Committee of ITQM 2014.

Keywords: CPI prediction; divide and conquer; ARIMA; gray model; BPNN.

* Corresponding author. Tel.: +86-10-82500904; fax: +86-10-62511184 .
E-mail address: weixu@ruc.edu.cn .

Selection and peer-review under responsibility of the Organizing Committee of ITQM 2014.
doi:10.1016/j.procs.2014.05.335

1. Introduction

The Consumer Price Index (CPI) is a widely used measurement of cost of living. It not only affects the government monetary, fiscal, consumption, prices, wages, social security, but also closely relates to the residents' daily life. As an indicator of inflation in China economy, the change of CPI undergoes intense scrutiny. For instance, The People's Bank of China raised the deposit reserve ratio in January, 2008 before the CPI of 2007 was announced, for it is estimated that the CPI in 2008 will increase significantly if no action is taken. Therefore, precisely forecasting the change of CPI is significant to many aspects of economics, some examples include fiscal policy, financial markets and productivity. Also, building a stable and accurate model to forecast the CPI will have great significance for the public, policymakers and research scholars.

Previous studies have already proposed many methods and models to predict economic time series or indexes such as CPI. Some previous studies make use of factors that influence the value of the index and forecast it by investigating the relationship between the data of those factors and the index. These forecasts are realized by models such as Vector autoregressive (VAR) model[1] and genetic algorithms-support vector machine (GA-SVM) [2].

However, these factor-based methods, although effective to some extent, simply rely on the correlation between the value of the index and limited number of exogenous variables (factors) and basically ignore the inherent rules of the variation of the time series. As a time series itself contains significant amount of information[3], often more than a limited number of factors can do, time series-based models are often more effective in the field of prediction than factor-based models.

Various time series models have been proposed to find the inherent rules of the variation in the series. Many researchers have applied different time series models to forecasting the CPI and other time series data. For example, the ARIMA model once served as a practical method in predicting the CPI[4]. It was also applied to predict submicron particle concentrations frommeteorological factors at a busy roadside in Hangzhou, China[5]. What's more, the ARIMA model was adopted to analyse the trend of pre-monsoon rainfall data forwestern India[6]. Besides the ARIMA model, other models such as the neural network, gray model are also widely used in the field of prediction. Hwang used the neural-network to forecast time series corresponding to ARMA (p, q) structures and found that the BPNNs generally perform well and consistently when a particular noise level is considered during the network training[7]. Aiken also used a neural network to predict the level of CPI and reached a high degree of accuracy[8]. Apart from the neural network models, a seasonal discrete grey forecasting model for fashion retailing was proposed and was found practical for fashion retail sales forecasting with short historical data and better than other state-of-art forecastingtechniques[9]. Similarly, a discrete Grey Correlation Model was also used in CPI prediction[10]. Also, Ma et al. used gray model optimized by particle swarm optimization algorithm to forecast iron ore import and consumption of China[11]. Furthermore, to deal with the nonlinear condition, a modified Radial Basis Function (RBF) was proposed by researchers[12].

In this paper, we propose a new method called "divide-and-conquer model" for the prediction of the CPI. We divide the total CPI into eight categories according to the CPI construction and then forecast the eight sub-CPIs using the GM (1, 1) model, the ARIMA model and the BPNN. To further improve the performance, we again make prediction of the sub-CPIs whose forecasting results are not satisfying enough by adopting new forecasting methods. After improvement and error adjustment, we get the advanced predicting results of the sub-CPIs. Finally we get the total CPI prediction by integrating the best forecasting results of each sub-CPI.

The rest of this paper is organized as follows. In section 2, we give a brief introduction of the three models mentioned above. And then the proposed model will be demonstrated in the section 3. In section 4 we provide the forecasting results of our model and in section 5 we make special improvement by adjusting the forecasting methods of sub-CPIs whose predicting results are not satisfying enough. And in section 6 we give elaborate discussion and evaluation of the proposed model. Finally, the conclusion is summarized in section 7.

2. Introduction to GM(1,1), ARIMA & BPNN

2.1. Introduction to GM(1,1)

The grey system theory is first presented by Deng in 1980s. In the grey forecasting model, the time series can be predicted accurately even with a small sample by directly estimating the interrelation of data. The GM(1,1) model is one type of the grey forecasting which is widely adopted. It is a differential equation model of which the order is 1 and the number of variable is 1, too. The differential equation is:

$$dx^{(1)}/_{dt} + ax^{(1)} \tag{1}$$

where $x^{(1)}$ is a sequence generated after accumulating, t is time, a and u are parameters to be estimated.

2.2. Introduction to ARIMA

Autoregressive Integrated Moving Average (ARIMA) model was first put forward by Box and Jenkins in 1970. The model has been very successful by taking full advantage of time series data in the past and present. ARIMA model is usually described as ARIMA (p, d, q), p refers to the order of the autoregressive variable, while d and q refer to integrated, and moving average parts of the model respectively. When one of the three parameters is zero, the model is changed to model "AR", "MR" or "ARMR". When none of the three parameters is zero, the model is given by:

$$(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - \mathrm{L})^d X_t = \delta + (1 + \sum_{i=1}^{q} \theta_i L^i) \tag{2}$$

where L is the lag number, ε_t is the error term.

2.3. Introduction to BPNN

Artificial Neural Network (ANN) is a mathematical and computational model which imitates the operation of neural networks of human brain. ANN consists of several layers of neurons. Neurons of contiguous layers are connected with each other. The values of connections between neurons are called "weight". Back Propagation Neural Network (BPNN) is one of the most widely employed neural network among various types of ANN. BPNN was put forward by Rumelhart and McClelland in 1985. It is a common supervised learning network well suited for prediction. BPNN consists of three parts including one input layer, several hidden layers and one output layer, as is demonstrated in Fig 1. The learning process of BPNN is modifying the weights of connections between neurons based on the deviation between the actual output and the target output until the overall error is in the acceptable range.

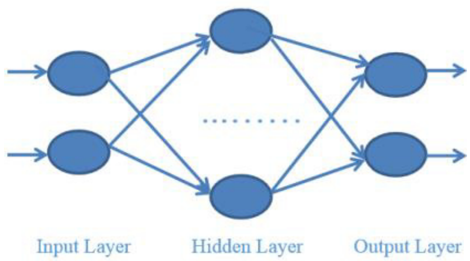

Fig. 1. Back-propagation Neural Network

3. The Proposed Method

3.1. The framework of the dividing-integration model

The process of forecasting national CPI using the dividing-integration model is demonstrated in Fig 2.

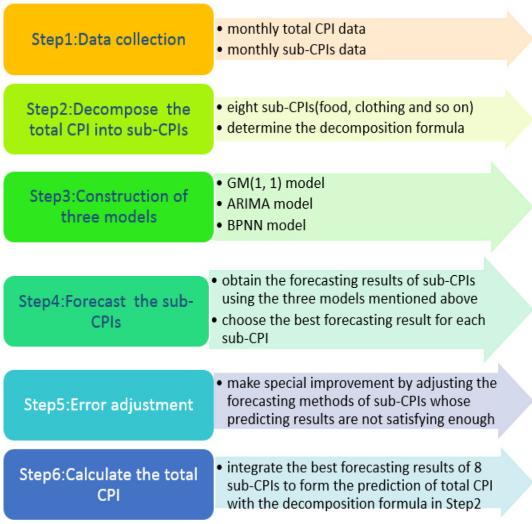

Fig. 2.The framework of the dividing-integration model

As can be seen from Fig. 2, the process of the proposed method can be divided into the following steps:

Step1: Data collection. The monthly CPI data including total CPI and eight sub-CPIs are collected from the official website of China's State Statistics Bureau (http://www.stats.gov.cn/).

Step2: Dividing the total CPI into eight sub-CPIs. In this step, the respective weight coefficient of eight sub-CPIs in forming the total CPI is decided by consulting authoritative source. (http://www.stats.gov.cn/). The eight sub-CPIs are as follows: 1. Food CPI; 2. Articles for Smoking and Drinking CPI; 3. Clothing CPI; 4. Household Facilities, Articles and Maintenance Services CPI; 5. Health Care and Personal Articles CPI; 6. Transportation and Communication CPI; 7. Recreation, Education and Culture Articles and Services CPI; 8. Residence CPI. The weight coefficient of each sub-CPI is shown in Table 8.

Table 1. 8 sub-CPIs weight coefficient in the total index

Index	No.1	No.2	No.3	No.4	No.5	No.6	No.7	No.8
Weight	31.79%	3.49%	8.52%	5.64%	9.64%	9.95%	13.75%	17.22%

Note: The index number stands for the corresponding type of sub-CPI mentioned before. Other indexes appearing in this paper in such form have the same meaning as this one.

So the decomposition formula is presented as follows:

$$TI = \sum_{i=1}^{8} W_i I_i \tag{3}$$

where TI is the total index; $I_i (i = 1, 2, \ldots, 8)$ are eight sub-CPIs. To verify the formula, we substitute historical numeric CPI and sub-CPI values obtained in Step1 into the formula and find the formula is accurate.

Step3: The construction of the GM (1, 1) model, the ARIMA (*p, d, q*) model and the BPNN model. The three models are established to predict the eight sub-CPIs respectively.

Step4: Forecasting the eight sub-CPIs using the three models mentioned in Step3 and choosing the best forecasting result for each sub-CPI based on the errors of the data obtained from the three models.

Step5: Making special improvement by adjusting the forecasting methods of sub-CPIs whose predicting results are not satisfying enough and get advanced predicting results of total CPI.

Step6: Integrating the best forecasting results of 8 sub-CPIs to form the prediction of total CPI with the decomposition formula in Step2.

In this way, the whole process of the prediction by the dividing-integration model is accomplished.

3.2. The construction of the GM(1,1) model

The process of GM (1, 1) model is represented in the following steps:

Step1: The original sequence: $x^{(0)} = \{x^{(0)}(1), x^{(0)}(2), x^{(0)}(3), \cdots, x^{(0)}(n)\}, i = 1, 2, \cdots, n$.

$$x^{(1)}(i) = \sum_{m=1}^{i} x^{(0)}(m), \quad i = 1, 2, \cdots, n \tag{4}$$

Step2: Estimate the parameters a and u using the ordinary least square (OLS).

Step3: Solve equation as follows.

$$\hat{x}^{(1)}(i+1) = (x^{(0)}(1) - \frac{u}{a})e^{-ai} + \frac{u}{a}, \qquad \begin{cases} \hat{x}^{(0)}(1) = \hat{x}^{(1)}(1) \\ \hat{x}^{(0)}(i) = \hat{x}^{(1)}(i) - \hat{x}^{(1)}(i-1), i = 2, 3, \cdots, n \end{cases} \tag{5}$$

Step4: Test the model using the variance ratio and small error possibility.

3.3. The construction of the ARIMA model

Firstly, ADF unit root test is used to test the stationarity of the time series. If the initial time series is not stationary, a differencing transformation of the data is necessary to make it stationary. Then the values of p and q are determined by observing the autocorrelation graph, partial correlation graph and the R-squared value.

After the model is built, additional judge should be done to guarantee that the residual error is white noise through hypothesis testing. Finally the model is used to forecast the future trend of the variable.

3.4. The construction of the BPNN model

The first thing is to decide the basic structure of BP neural network. After experiments, we consider 3 input nodes and 1 output nodes to be the best for the BPNN model. This means we use the CPI data of timet_0, t_1, t_2 toforecast the CPI of timet_3.

The hidden layer level and the number of hidden neurons should also be defined. Since the single-hidden-layer BPNN are very good at non-liner mapping, the model is adopted in this paper. Based on the Kolmogorov theorem and testing results, we define 5 to be the best number of hidden neurons. Thus the 3-5-1 BPNN structure is determined.

As for transferring function and training algorithm, we select 'tansig' as the transferring function for middle layer, 'logsig' for input layer and 'traingd' as training algorithm. The selection is based on the actual performance of these functions, as there are no existing standards to decide which ones are definitely better than others.

Eventually, we decide the training times to be 35000 and the goal or the acceptable error to be 0.01.

4. Empirical Analysis

CPI data from Jan. 2012 to Mar. 2013 are used to build the three models and the data from Apr. 2013 to Sept. 2013 are used to test the accuracy and stability of these models. What's more, the MAPE is adopted to evaluate the performance of models. The MAPE is calculated by the equation:

$$MAPE = 1/n\sum_{i=1}^{n}\left(\left|predicting\ value - real\ value\right|/real\ value\right) \tag{6}$$

4.1. Data source

An appropriate empirical analysis based on the above discussion can be performed using suitably disaggregated data. We collect the monthly data of sub-CPIs from the website of National Bureau of Statistics of China (http://www.stats.gov.cn/).

Particularly, sub-CPI data from Jan. 2012 to Mar. 2013 are used to build the three models and the data from Apr. 2013 to Sept. 2013 are used to test the accuracy and stability of these models.

4.2. Experimental results

We use MATLAB to build the GM (1,1) model and the BPNN model, and Eviews 6.0 to build the ARIMA model. The relative predicting errors of sub-CPIs are shown in Table 2.

Table 2.Error of Sub-CPIs of the 3 Models

	No.1	No.2	No.3	No.4	No.5	No.6	No.7	No.8
GM(1,1)	0.0378	0.0009	0.0075	0.0031	0.0010	0.0047	0.0061	0.0047
ARIMA	0.0059	0.0025	0.0027	0.0016	0.0010	0.0072	0.0026	0.0016
BPNN	0.0110	0.0193	0.0041	0.0036	0.0058	0.0053	0.0029	0.0016

From the table above, we find that the performance of different models varies a lot, because the characteristic of the sub-CPIs are different. Some sub-CPIs like the Food CPI changes drastically with time while some do not have much fluctuation, like the Clothing CPI. We use different models to predict the sub-CPIs and combine them by equation 7.

$$\mathrm{Y} = \sum_{i=1}^{8} c_i x_i \tag{7}$$

Where Y refers to the predicted rate of the total CPI, c_i is the weight of the sub-CPI which has already been shown in Table 1andx_i is the predicted value of the sub-CPI which has the minimum error among the three models mentioned above. The model chosen will be demonstrated in Table 3:

Table 3.The model used to forecast

	No.1	No.2	No.3	No.4	No.5	No.6	No.7	No.8
Model	ARIMA	GM(1,1)	ARIMA	ARIMA	GM(1,1)	GM(1,1)	ARIMA	BPNN
Error	0.0059	0.0009	0.0028	0.0016	0.0010	0.0047	0.0026	0.0016

After calculating, the error of the total CPI forecasting by the dividing-integration model is 0.0034.

5. Model Improvement & Error Adjustment

As we can see from Table 3, the prediction errors of sub-CPIs are mostly below 0.004 except for two sub-CPIs: Food CPI whose error reaches 0.0059 and Transportation & Communication CPI 0.0047.

In order to further improve our forecasting results, we modify the prediction errors of the two aforementioned sub-CPIs by adopting other forecasting methods or models to predict them. The specific methods are as follows.

5.1. Error adjustment of food CPI

In previous prediction, we predict the Food CPI using the BPNN model directly. However, the BPNN model is not sensitive enough to investigate the variation in the values of the data. For instance, although the Food CPI varies a lot from month to month, the forecasting values of it are nearly all around 103.5, which fails to make meaningful prediction.

We ascribe this problem to the feature of the training data. As we can see from the original sub-CPI data on the website of National Bureau of Statistics of China, nearly all values of sub-CPIs are around 100. As for Food CPI, although it does have more absolute variations than others, its changes are still very small relative to the large magnitude of the data (100). Thus it will be more difficult for the BPNN model to detect the rules of variations in training data and the forecasting results are marred.

Therefore, we use the first-order difference series of Food CPI instead of the original series to magnify the relative variation of the series forecasted by the BPNN. The training data and testing data are the same as that in previous prediction. The parameters and functions of BPNN are automatically decided by the software, SPSS.

We make 100 tests and find the average forecasting error of Food CPI by this method is 0.0028. The part of the forecasting errors in our tests is shown as follows in Table 4:

Table 4.The forecasting errors in BPNN test

No.	1	2	3	4	5	6	7	8	9	10
Error	0.2555	0.2509	0.3329	0.1841	0.2659	0.2838	0.3341	0.2679	0.2851	0.2886

No.	11	12	13	14	15	16	17	18	19	20
Error	0.3159	0.2766	0.3132	0.2894	0.3267	0.3197	0.332	0.252	0.333	0.3068

5.2. Error adjustment of transportation &communication CPI

We use the Moving Average (MA) model to make new prediction of the Transportation and Communication CPI because the curve of the series is quite smooth with only a few fluctuations.

We have the following equation(s):

$$S_t = \alpha x_t + (1 - \alpha) S_{t-1} \tag{8}$$

$$\hat{X}_{t+1} = S_t \tag{9}$$

where $x_1, x_2 \dots x_n$ is the time series of the Transportation and Communication CPI, S_t is the value of moving average at time t, α is a free parameter which should be decided through experiment.

To get the optimal model, we range the value of α from 0 to 1. Finally we find that when the value of α is 0.95, the forecasting error is the smallest, which is 0.0039.

The predicting outcomes are shown as follows in Table5:

Table 5.The Predicting Outcomes of MA model

Date	Actual value	Predicted value	Residual error	Relative error
2013-4	98.9	99.7098	-0.8098	0.0081
2013-5	98.8	98.9162	-0.1162	0.0012
2013-6	99.3	98.8023	0.4977	0.0050
2013-7	99.9	99.2901	0.6099	0.0061
2013-8	100	99.8878	0.1122	0.0011
2013-9	99.8	99.9978	-0.1978	0.0020

5.3. Advanced results after adjustment to the models

After making some adjustment to our previous model, we obtain the advanced results as follows in Table 6:

Table 6.The model used to forecast and the Relative Error

	No.1	No.2	No.3	No.4	No.5	No.6	No.7	No.8
Model	BPNN	GM(1,1)	ARIMA	ARIMA	GM(1,1)	MA	ARIMA	BPNN
Error	0.0028	0.0009	0.0027	0.0016	0.0010	0.0039	0.0026	0.0016

After calculating, the error of the total CPI forecasting by the dividing-integration model is 0.2359.

6. Further Discussion

To validate the dividing-integration model proposed in this paper, we compare the results of our model with the forecasting results of models that do not adopt the dividing-integration method. For instance, we use the ARIMA model, the GM (1, 1) model, the SARIMA model, the BRF neural network (BRFNN) model, the Verhulst model and the Vector Autoregression (VAR) model respectively to forecast the total CPI directly without the process of decomposition and integration. The forecasting results are shown as follows in Table7.

Table 7.The prediction error of other models

Model	SARIMA	BRF NN	Verhulst	ARIMA	GM(1,1)	VAR
Error	0.0038	0.0057	0.0035	0.0038	0.0090	0.0179

From Table 7, we come to the conclusion that the introduction of dividing-integration method enhances the accuracy of prediction to a great extent. The results of model comparison indicate that the proposed method is not only novel but also valid and effective.

The strengths of the proposed forecasting model are obvious. Every sub-CPI time series have different fluctuation characteristics. Some are relatively volatile and have sharp fluctuations such as the Food CPI while others are relatively gentle and quiet such as the Clothing CPI. As a result, by dividing the total CPI into several sub-CPIs, we are able to make use of the characteristics of each sub-CPI series and choose the best forecasting model among several models for every sub-CPI's prediction. Moreover, the overall prediction error is provided in the following formula:

$$TE = \sum_{i=1}^{8} c_i e_i \tag{10}$$

where *TE* refers to the overall prediction error of the total CPI, c_i is the weight of the sub-CPI shown in table 1 and e_i is the forecasting error of corresponding sub-CPI.

In conclusion, the dividing-integration model aims at minimizing the overall prediction errors by minimizing the forecasting errors of sub-CPIs.

7. Conclusions and future work

This paper creatively transforms the forecasting of national CPI into the forecasting of 8 sub-CPIs. In the prediction of 8 sub-CPIs, we adopt three widely used models: the GM (1, 1) model, the ARIMA model and the BPNN model. Thus we can obtain the best forecasting results for each sub-CPI. Furthermore, we make special improvement by adjusting the forecasting methods of sub-CPIs whose predicting results are not satisfying enough and get the advanced predicting results of them. Finally, the advanced predicting results of the 8 sub-CPIs are integrated to form the forecasting results of the total CPI.

Furthermore, the proposed method also has several weaknesses and needs improving. Firstly, The proposed model only uses the information of the CPI time series itself. If the model can make use of other information such as the information provided by factors which make great impact on the fluctuation of sub-CPIs, we have every reason to believe that the accuracy and stability of the model can be enhanced. For instance, the price of pork is a major factor in shaping the Food CPI. If this factor is taken into consideration in the prediction of Food CPI, the forecasting results will probably be improved to a great extent. Second, since these models forecast the future by looking at the past, they are not able to sense the sudden or recent change of the environment. So if the model can take web news or quick public reactions with account, it will react much faster to sudden incidence and affairs. Finally, the performance of sub-CPIs prediction can be higher. In this paper we use GM (1, 1), ARIMA and BPNN to forecast sub-CPIs. Some new method for prediction can be used. For instance, besides BPNN, there are other neural networks like genetic algorithm neural network (GANN) and wavelet neural network (WNN), which might have better performance in prediction of sub-CPIs. Other methods such as the VAR model and the SARIMA model should also be taken into consideration so as to enhance the accuracy of prediction.

Acknowledgement

This research work was partly supported by 973 Project (Grant No. 2012CB316205), National Natural Science Foundation of China (Grant No.71001103, 91224008, 91324015), Beijing Social Science Fund (No. 13JGB035), Beijing Natural Science Foundation (No. 9122013), Beijing Nova Program (No.Z131101000413058), and Program for Excellent Talents in Beijing.

References

1. Wang W, Wang T, and Shi Y. Factor analysis on consumer price index rising in China from 2005 to 2008. *Management and service science* 2009; p. 1-4.
2. Qin F, Ma T, and Wang J. The CPI forecast based on GA-SVM. *Information networking and automation* 2010; p. 142-147.
3. George EPB, Gwilym MJ, and Gregory CR. *Time series analysis: forecasting and control*. 4th ed. Canada: Wiley; 2008
4. Weng D. The consumer price index forecast based on ARIMA model. *WASE International conferenceon information engineering* 2010; p. 307-310.
5. Jian L, Zhao Y, Zhu YP, Zhang MB, Bertolatti D. An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. *Science of total enviroment* 2012;**426**:336-345.
6. Priya N, Ashoke B, Sumana S, Kamna S. Trend analysis and ARIMA modelling of pre-monsoon rainfall data forwestern India. *Comptes rendus geoscience* 2013;**345**:22-27.
7. Hwang HB. Insights into neural-network forecasting of time seriescorresponding to ARMA(p; q) structures. *Omega* 2001;**29**:273-289.
8. Milam A. Using a neural network to forecast inflation. *Industrial management & data systems* 1999;**7**:296-301.
9. Min X, Wong WK. A seasonal discrete grey forecasting model for fashion retailing. *Knowledge based systems* 2014;**57**:119-126.
11. Weimin M, Xiaoxi Z, Miaomiao W. Forecasting iron ore import and consumption of China using grey model optimized by particle swarm optimization algorithm. *Resources policy* 2013;**38**:613-620.
12. Zhen D, and Feng S. A novel DGM (1, 1) model for consumer price index forecasting. *Grey systems and intelligent services (GSIS)* 2009; p. 303-307.
13. Yu W, and Xu D. Prediction and analysis of Chinese CPI based on RBF neural network. *Information technology and applications* 2009;**3**:530-533.
14. Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* 2003;**50**:159-175.
15. Pai PF, Lin CS. A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega* 2005;**33(6)**:497-505.
16. Tseng FM, Yu HC, Tzeng GH. Combining neural network model with seasonal time series ARIMA model. *Technological forecasting and social change* 2002;**69(1)**:71-87.
17. Cho MY, Hwang JC, Chen CS. Customer short term load forecasting by using ARIMA transfer function model. *Energy management and power delivery, proceedings of EMPD'95. 1995 international conference on IEEE*, 1995;**1**:317-322.