Application of Time Sequence Model Based on Excluded Seasonality in Daily Runoff Prediction

She Wei[1,2,*] and Li Difang[1]

[1]School of Mathematics and Statistics, Wuhan University, Wuhan, 430072, China; [2]School of Mathematics and Statistics, South-central University for Nationalities, Wuhan, 430074, China

Abstract: To build a time sequence prediction model with excluded seasonality out of the sequence with prominent seasonal features, primary treatment of seasonality exclusion shall firstly be done, to obtain the pass of stationary test; and then, by integrating autocorrelogram, partial autocorrelogram and AIC codes, ARMA model shall be identified and be able to pass residual correlation test and the optimum hydrological probability distribution function shall be finally determined. In accordance with the daily runoff of the Three Gorges from 1950-2009, the predicted daily runoff in 2010 *via* R software has small error compared with the observed daily runoff, obtaining good prediction performance.

Keywords: AIC codes, daily runoff prediction, seasonality exclusion, time sequence prediction.

1. INTRODUCTION

In recent years, the frequency analysis method for non-uniformity hydrological sequence under varying conditions has been studied, in hope of successful analysis on the evolvement mechanism of hydrological statistics discipline under varying conditions, from the hydrological features that are under variation [1] out of climate change and human activities. As a part of hydrological frequency analysis, daily runoff prediction bears practical significance in the micro-management, optimized dispatching, development and utilization as well as flood control and drought relief decision of regional water resource. As for the current study on models for daily runoff sequence prediction, there are such models available as seasonal auto regression model, staged smooth auto regression model, artificial neural network model [2], and maximum likelihood method and least square method models for non-stationary hydrological frequency analysis.

Therein, the seasonal auto regression model features simple principles and convenient computing; however, specifically for the daily runoff of the Three Gorges, there are 365 sets of parameters to be considered, thus leading to lengthy and complex computing [3]. The staged smooth auto regression model, despite of its explicit concept and simple structure, bears big error in predicting certain mutational sites of daily runoff sequence curve. As for artificial neural network model, there is certain motility in parameter determination and no defined algorithm, nullifying the practical significance of parameters [4]. For the treatment of non-stationary hydrological elements (tendency elements) in Flood Frequency Model (FHM), Strupczewski *et al.* put forward the maximum likelihood method and least square method models for non-stationary hydrological frequency analysis and successfully applied the two methods in the hydrological analysis on Polish River [5-7]. This thesis, based on work of Strupczewski *et al.*, observed the tendency of the mean value and variance for daily runoff sequence of the Three Gorges that varies from time, found out the reason for varying hydrological regime of the Three Gorges, adopted seasonality exclusion method to modify non-uniformity and maximum likelihood method to estimate the tendency function and hydrological distribution parameters, and integrated autocorrelogram, partial autocorrelogram and AIC codes to identify ARMA model and finally determine the optimum hydrological probability distribution.

2. BUILDING OF ARMA MODEL

2.1. Seasonality Exclusion of Sequence

In general, the hydrological sequence features tendency, periodicity and seasonality, among which the strong seasonality is the main reason for the non-linearity of the whole sequence, which will influence the accuracy of time sequence analysis and thus require seasonality exclusion. Normally, there are three methods for seasonality exclusion [8]: Seasonality Auto regression Integrated Moving Average (SARIMA) model, ARMA model with excluded seasonal factors and periodical ARMA model. Based on the variation features of hydrological sequence, ARMA model is used in this thesis for seasonality exclusion, as shown below:

$$u_{r.m} = \frac{x_{r.m} - \mu_m}{\sigma_m} \tag{1}$$

Therein, r means the year, m means the certain time within a year (in case of daily data, m may be maximally taken as 366; in case of monthly data, m is taken as 12 for maximum), $x_{r.m}$ means the original sequential value at time point m in year r, μ_m means the average value at time point m in each year, σ_m means the standard deviation at time point m

in each year, and $\mu_{r.m}$ means the sequential value at time point m in year r after seasonality exclusion.

The above seasonality exclusion method is similar to the evaluation of variable coefficient, which is the characteristic number to measure the evaluation fluctuation degree of random variables in mathematical expectation, thus being a dimensionless quantity, the influence exclusion of which can indirectly lead to the exclusion of seasonality of daily runoff sequence to make it more stationary.

2.2. Discrimination of Sequence Correlation

Usually, there are three approaches for sequence correlation test [8]: autocorrelation coefficient, partial autocorrelation coefficient and Ljung-Box Q statistics, among which the former two are used for identifying the ARMA model order, while the latter is used for testing the correlation of sequential residual.

Assume that u_t means the treated sequence, its autocorrelation coefficient of the lagged order k shall be estimated in the following formula:

$$r_k = \frac{\sum_{t=k+1}^{T}\left(u_t - \bar{u}\right)\left(u_{t-k} - \bar{u}\right)}{\sum_{t=1}^{T}\left(u_t - \bar{u}\right)^2} \tag{2}$$

Therein, $\bar{u}$ is the sample average of the sequence; r_k is stated as the autocorrelation coefficient of order k for time sequence u_t, and such coefficient can represent the partial correlation among adjacent data of sequence u_t. Partial autocorrelation coefficient means the conditional correlation between u_t and u_{t-1} under the given u_{t-1}, u_{t-2} and Lu_{t-k-1}. The degree of correlation is measured by partial autocorrelation coefficient $\varphi_{k,k}$, the estimation of which under lagged order k is shown in the following calculation formula:

$$\varphi_{k,k} = \begin{cases} r_1, k = 1 \\ \dfrac{r_k - \sum_{j=1}^{k-1}\varphi_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1}\varphi_{k-1,j} r_{k-j}}, k > 1 \end{cases} \tag{3}$$

$$\varphi_{k,j} = \varphi_{k-1,j} + \varphi_{k,k}\varphi_{k-1,k-j}, \left(k \neq j\right).$$

The expression of Ljung-Box Q statistics is as follows:

$$Q_{LB} = T\left(T+2\right)\sum_{j=1}^{p}\frac{r_j^2}{T-j} \tag{4}$$

Therein, r_j is the autocorrelation coefficient of order j, T is the sample size, and p is the set lagged order. The original assumption of Q_{LB} statistics is that no autocorrelation of order p for the sequence exists, instead, such sequence subjects to the x^2 distribution. Usually, Q_{LB} statistics requires large sample size to ensure its validity.

2.3. Identification of ARMA Model

The stationary ARMA model of the treated sequence is generally composed of auto regression model and moving average model and marked as ARMA (p, q), wherein p means the maximum order of autoregressive process, and q means the maximum order of moving average. The expression of ARMA (p, q) is as follows:

$$u_t = c + \phi_1 u_{t-1} + \phi u_{t-2} + \cdots + \phi_p u_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-p} \tag{5}$$

Therein, c is the constant, $\varphi_1,\ldots,\varphi_p$ are parameters of auto regression model, p is the order of auto regression model, $\theta_1,\ldots,\theta_q$ are parameters of moving average model, q is the order of moving average model, and ε_t is the white noise sequence with the mean value of 0 and variance of definite value.

Generally, during the identification of ARMA model, Autocorrelogram (ACF) and Partial Autocorrelogram (PACF) of the sequence are observed, and the model order is determined by trailing or truncation. The specific decision rules are shown in Table **1**.

Generally, the models can be well identified through the above methods; however, the residual correlation test will always not be passed when the characteristics of original sequence is too complex. Such problem can be solved through integrated decision method, and the specific steps are as follows:

(1) By observing partial autocorrelogram, the maximum orders p and q, which are respectively greater than 2 times of positive and negative standard deviation [9], are determined.

Calculate AIC_{p_1,q_1}, wherein $p_1 \in (1,p), q_1 \in (1,q)$.

(2) Make the value of AIC_{p_1,q_1} to be minimum, and the corresponding p and q as the orders for model identification. Therein, AIC refers to the Akaike Information Criterion [9, 10], with the calculation formula being:

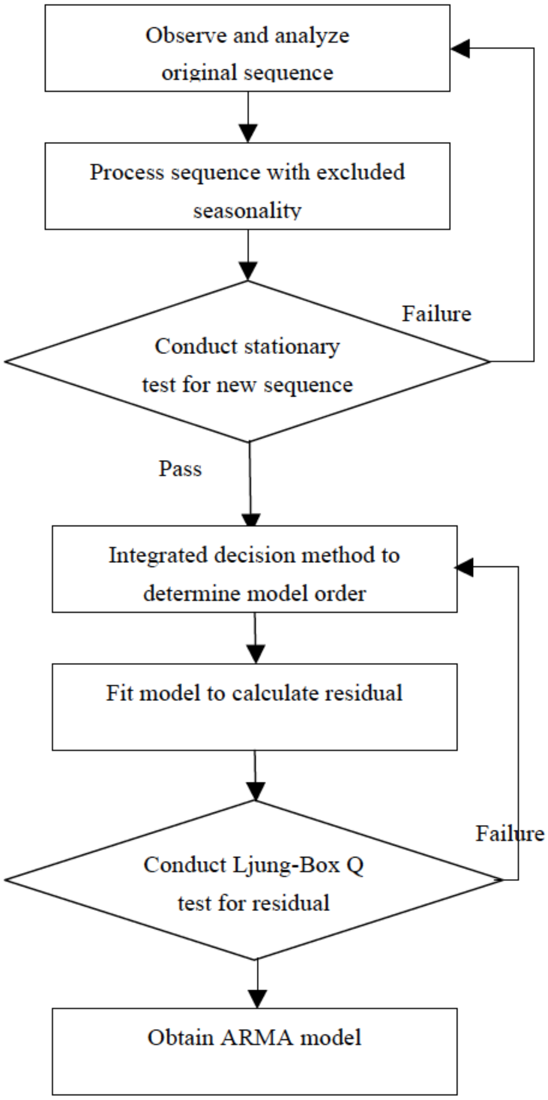
$$AIC = -2\log(\mathrm{ML}) + 2k \tag{6}$$

If the model involves an intercept or a constant term, then $k = p + q + 1$; or else, $k = p + q$.

Table 1. Rules of ARMA model decision.

	AR(p)	MA(q)	ARMA(p, q)
ACF	Trailing	Truncation after lagging for order q	Trailing
PACF	Truncation after lagging for order p	Trailing	Trailing

2.4. Calculation Process


Observe and analyze original sequence
Process sequence with excluded seasonality
Conduct stationary test for new sequence
Failure
Pass
Integrated decision method to determine model order
Fit model to calculate residual
Conduct Ljung-Box Q test for residual
Failure
Obtain ARMA model

3. SOLUTION OF MODEL

Yichang Hydrologic Station, located in Yichang City, Hubei Province, China with controlling drainage area of 1,005,500 km^2, is an outlet control station in the upper reaches of the Yangtze River, with the perennial average runoff volume (1950-2010) being 13634m^3/s. The daily runoff data from 1950 to 2010 at the Station is selected in this thesis for time sequence analysis and analog prediction, wherein the data from 1950 to 2009 will be used for modeling and the data of 2010 will be used for testing model precision.

3.1. Preliminary Analysis on Sequence

The daily runoff data from 1950 to 2010 at Yichang Hydrologic Station is as shown in Fig. (**1**):

It can be seen from Fig. (**1**) that the sequence bears strong seasonality and long periodicity. In order to further master the discipline of runoff sequence, perennial daily mean value and daily standard deviation shall be calculated, with the result as shown in Fig. (**2**).

According to Fig. (**2**), the annual runoff sequence can be divided into four seasons, namely dry season prior to flood (January-March), transitional season prior to flood (April-May), flood season (June-October) and dry season after flood (November-December). The difference in daily mean values of each season shows the tendency of sequence, which is consistent with that of the seasonality of sequence.

3.2. Sequence Stationary Test

Since the sequence is generally required to be stationary when building time sequence model, while practically, it can be seen from the preliminary analysis of sequence that the runoff sequence is seasonal, periodical, non-stationary and of certain tendency, it's better to make the sequence as stationary as possible before building time sequence, to ensure the effect of fitting.

3.2.1. Preliminary Treatment for Sequence

Specific to the tendency, seasonality and periodicity of runoff sequence, we herein conduct seasonality exclusion for the original runoff sequence, the result of which is as shown in Fig. (**3**).

It can be preliminarily judged that the treated runoff sequence is almost stationary from Fig. (**3**); however, ADF test and PP test are still required for the treated sequence. The original assumption of the two tests is that the sequence is non-stationary. The test results are shown in Table **2**.

It can be seen from Table **2**. That the tests accept the alternative assumption instead of the original one, *i.e.* the sequence may be deemed as stationary at 95% confidence coefficient and thus ARMA model can be built.

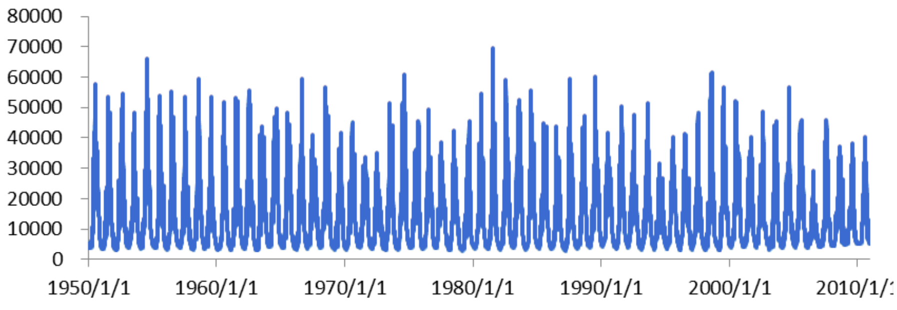


Fig. (1). Daily runoff at yichang hydrologic station in 1950-2010.

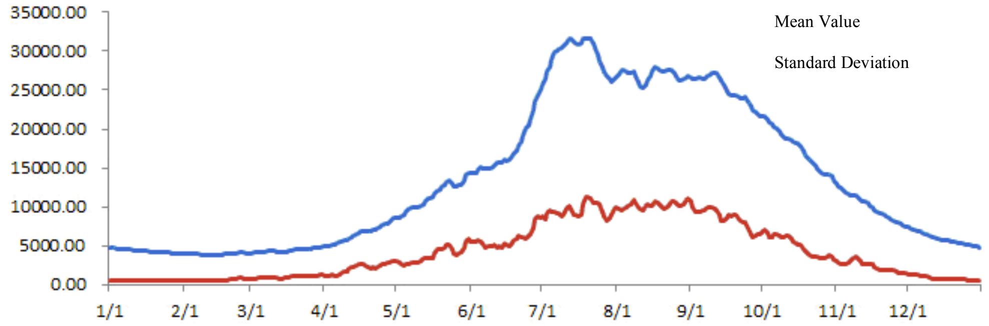


Fig. (2). Changes of mean value and standard deviation of daily runoff at yichang hydrologic station.

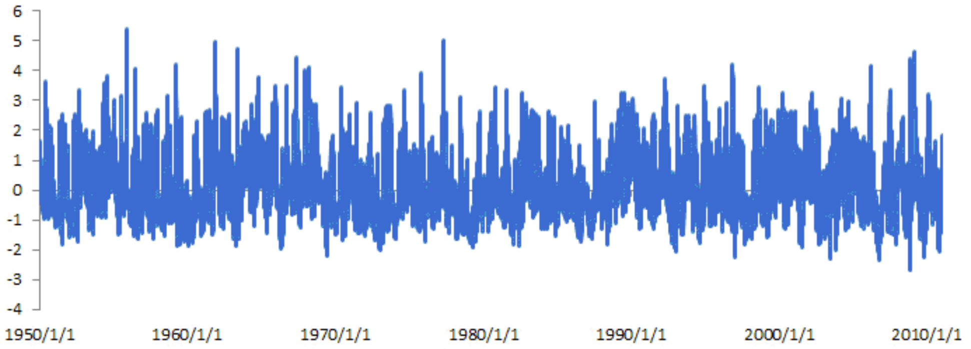


Fig. (3). The sequence after exclusion of seasonality.

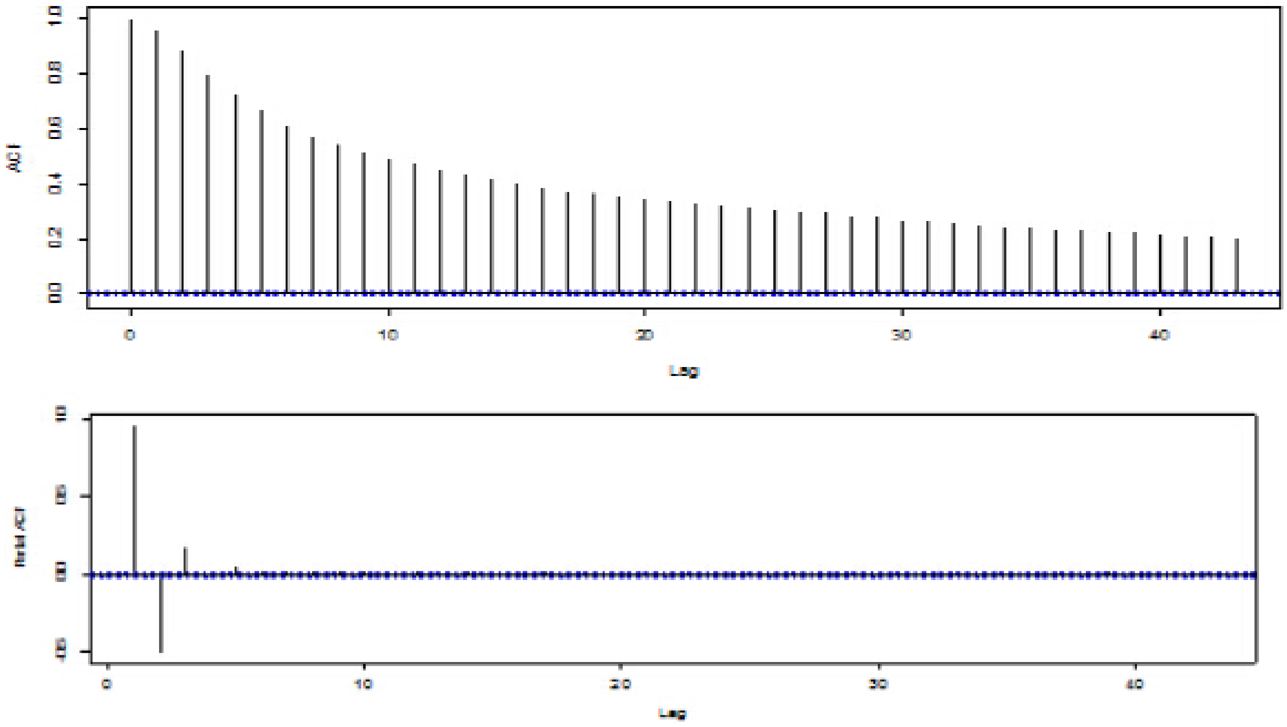


Fig. (4). Autocorrelogram and partial autocorrelogram of sequence after exclusion of seasonality.

Table 2. Stationary test of sequence after exclusion of seasonality.

Test Method	Statistics	p-Value
ADF test	-16.5507	0.01
PP test	-1035.042	0.01

3.2.2. Identification and Solution of Model

The identification of ARMA model involves the observation of autocorrelogram and partial autocorrelogram, AIC codes and BIC codes. It can be seen from Fig. (**4**) that within 95% confidence interval, the autocorrelation coefficients are in trailing, while partial autocorrelation coefficients are prominent before the first 10 orders except the order 4, *i.e.*

Table 3. Test the P value and AIC value by Ljung-Box Q test under different AR orders.

AR Order	P Value	>0.05	AIC Value	Compare with Last Order
3	9.659E-15	No	-1076.21	
4	2.24E-14	No	-1074.56	raise
5	1.12E-04	No	-1147.00	decline
6	5.98E-04	No	-1158.30	decline
7	2.01E-03	No	-1166.65	decline
8	8.51E-03	No	-1176.04	decline
9	0.0937	Yes	-1189.79	decline
10	0.153	Yes	-1194.22	decline
11	0.142	Yes	-1193.62	raise
12	0.375	Yes	-1199.40	raise

order 3 is of truncation. If Fig. (**4**) is the only basis, the AR (3) model may be built for the treated runoff sequence data from 1950 to 2009. When conducting residual correlation test for such model, the Ljung-Box Q statistics $P = 9.659 \times 10^{-15} < 0.01$, the original assumption is denied, namely, autocorrelation exists in residual sequence.

The ARMA model obtained only by observing autocorrelogram and partial autocorrelogram will not pass residual correlation test. To better identify the model to make it pass such test, autocorrelogram, partial autocorrelogram and a series of codes are integrated in this thesis to identify the model order, ensuring the ARMA model passes the test.

It can be seen from Table **3** that when the AR order is 10, the corresponding Ljung-Box Q test value *P*=0.153, being relatively prominent; the AIC value before order 11 is minimum, thus meeting AIC minimization code [7]. Therefore, the order of AR model is taken as *n*=10. Assume that u_t means the treated sequence; the following can be obtained through R software programming:

$$u_t = 1.5352u_{t-1} - 0.7834u_{t-2} + 0.2309u_{t-3} - 0.08u_{t-4} + \\ 0.0354u_{t-5} + 0.0032u_{t-6} + 0.0041u_{t-7} - 0.005u_{t-8} + \\ 0.003u_{t-9} + 0.0171u_{t-10} \tag{7}$$

As the coefficients of $u_{t-6}, u_{t-7}, u_{t-8}, u_{t-9}$ herein are not prominent, these coefficients shall be taken as 0. New relation can be obtained under re-fitting:

$$u_t = 1.5343u_{t-1} - 0.7818u_{t-2} + 0.2305u_{t-3} - 0.0808u_{t-4} + \\ 0.0399u_{t-5} + 0.0163u_{t-10} \tag{8}$$

3.2.3. Model Test

The traditional time sequence model also requires no residual correlation, so Ljung-Box Q test shall be conducted. It can be seen from Fig. (**5**) that:

① The residual autocorrelation coefficients are all within 95% confidence interval, *i.e.* the residual bears no prominent autocorrelation;

② The probability values (i.e p value) of Ljung-Box Q statistics are all beyond 95% confidence interval, and the original assumption is accepted, namely, the residual sequence bears no autocorrelation.

The conclusion that the residual sequence bears no autocorrelation can be drawn from ① and ②, further proving the validity of the model.

3.2.4. Prediction of Runoff Sequence

Prediction is an important application of time sequence. The daily runoff volume value of the Three Gorges in 2010 is herein obtained by multi-step prediction [8] taking the runoff from 1950-2009 at Yichang Hydrologic Station as original data with AR (10) model.

It can be seen from Fig. (**6**) that the predicted results and observed results are generally matching. The fluctuation of predicted results also reflects seasonality, meaning that the AR (10) built can well fit the runoff sequence; from Fig. (**7**) we can see that the bound of predicted value is always an interval belt with the same width, which depends on the assumption of AR (10) model, *i.e.* assume that the residual variance is a definite value, and the 95% confidence interval is formed from predicted value ±1.96 times of standard deviation. Therefore, the width of predicted confidence interval will be a definite value when the variance is constant.

CONCLUSION

By observing the autocorrelogram and partial autocorrelogram, general time sequence model can be well identified. However, featured by periodicity, tendency and seasonality, the daily runoff sequence of the Three Gorges has invalidated general model identification methods. Therefore, seasonality exclusion is done for the daily runoff sequence to make it more stationary.

After that, ARMA model is identified by integrating autocorrelogram, partial autocorrelogram and AIC codes. According to the result of Ljung-Box Q test, such model after identification can well fit the daily runoff sequence of the Three Gorges, with small error in prediction of daily

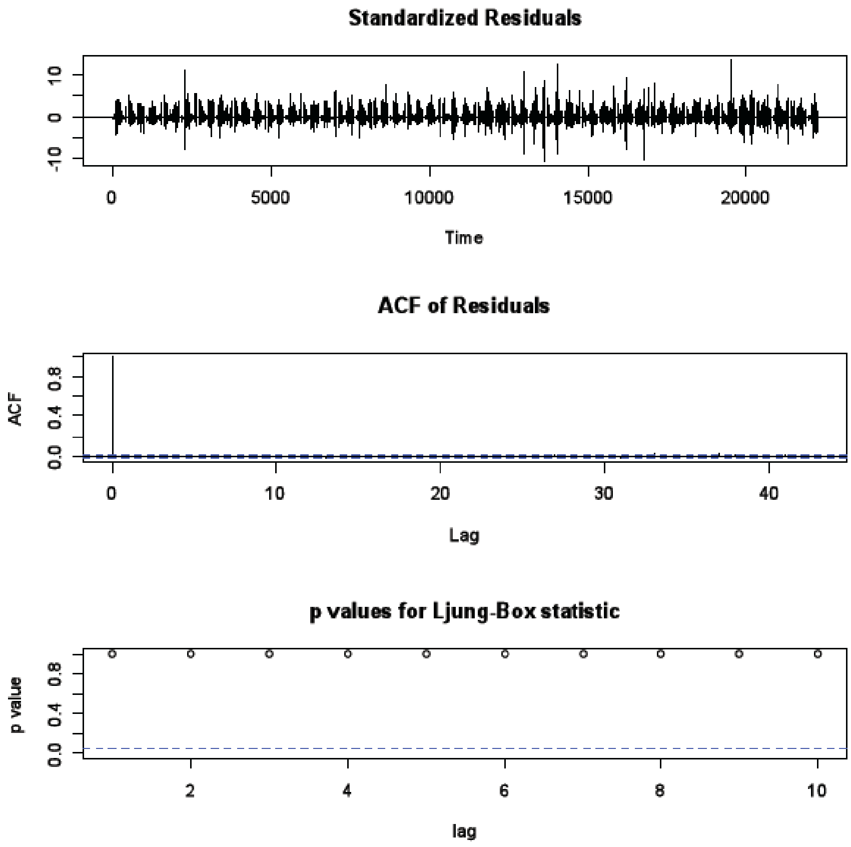


Fig. (5). Residual correlation test.

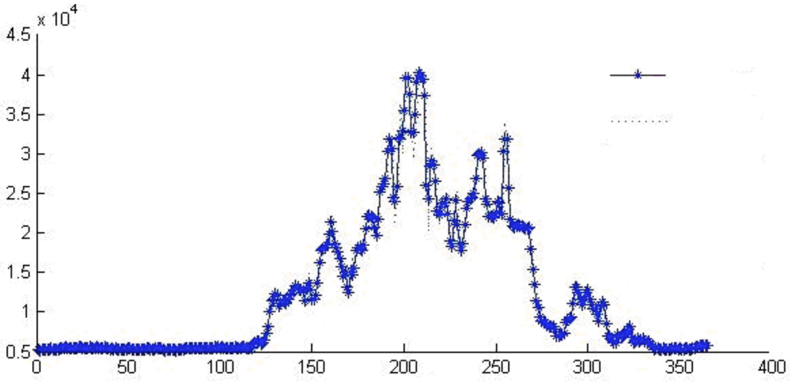


Fig. (6). Comparison of predicted results and observed results of *AR* (10) model.

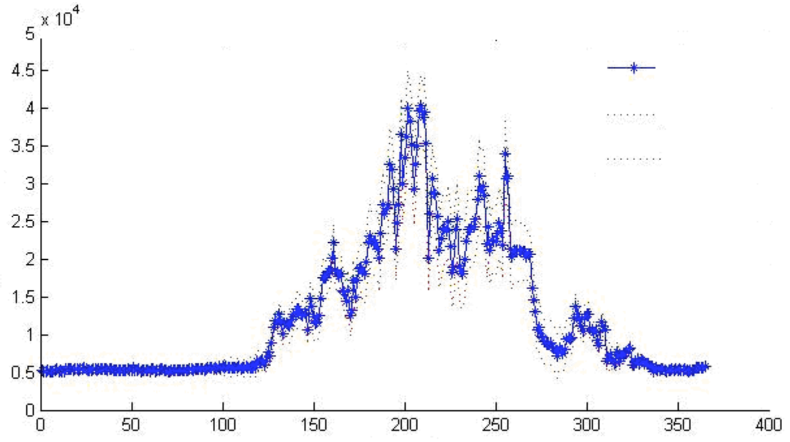


Fig. (7). Predict Results in 95%-confidence interval by *AR* (10) model.

runoff sequence value of 2010 and better predictive effects. The prediction method in this thesis can well solve such similar problems in sequence prediction under prominent seasonality.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by the Key Project of National Natural Sciences, China (No.51190095).

REFERENCES

[1] "*Compiling Committee of National Assessment Report of Climate Change*," National Assessment Report of Climate Change. Science Press, Beijing, 2007.

[2] C. Zhu, *Study on Medium-and-long-term Hydrologic Prediction of Runoff*, Sichuan University, Sichuan, 2005.

[3] W.G. Strupczewski, V.P. Singh, and W. Feluch, "Non-stationary approach to at-site flood frequency modeling I.Maximum likelihood estimation," *Journal of Hydrology,* vol. 248, no. 1-4, pp. 123-142, 2001.

[4] W.G. Strupczewski, and Z. Kaczmarek, "Non-stationary approach to at-site flood frequency modeling II. Weighted least squares estimation," *Journal of Hydrology*, vol. 248, no. 1-4, pp. 143-151, 2001.

[5] W.G. Strupczewski, V.P. Singh, and H.T. Mitosek, "Non-stationary approach to at-site flood frequency modeling III.Flood analysis of Polish rivers," *Journal of Hydrology*, vol. 248, no. 1-4, pp. 152-167, 2001.

[6] H. Wang, X. Gao, L. Qian, and S. Yu, "The Uncertainty Analysis of Hydrologic Process Based on ARMA-GARCH Model," *Science China: Technical Science*, vol. 42 no. 9, pp. 1069-1080, 2012.

[7] C.D. Jonathan, and K.S. Chan, *Time Series Analysis and Its Applications*, Beijing, China Machine Press, 2013.

[8] R.S. Tsay, *An Introduction to Analysis of Financial Data with R*, Beijing, China Machine Press, pp. 28-63, 2013.

[9] H. Dehling, A. Rooch, and M.S. Taqqu "Non-parametric change point tests for long-range dependent data," *Scandinavian Journal of Statistics*, vol. 40, no. 1, pp. 153-173, 2012.

[10] S. Li, L. Xiong, L. Dong, and J. Zhang, "Effects of the Three Gorges Reservoir on the hydrological droughts at the downstream Yichang station during 2003-2011," *Hydrological Processes,* vol. 27, no. 26, pp. 3981-3993, 2013.

Received: September 22, 2014 Revised: November 30, 2014 Accepted: December 02, 2014