

基于 ARMA 模型的水文数据预测

陈佳袁¹, 闫杰²

(1. 桐乡市水利局, 浙江 桐乡 314500;
2. 中国电力工程顾问集团西北电力设计院有限公司, 陕西 西安 710075)

摘要: 针对小样本水文数据序列难以准确预测的特点, 将时间序列分析运用于水文数据的预测分析, 研究基于 AIC 定阶准则和遗传算法定阶的 ARMA 模型, 并将其运用于周期性水文数据的预测。根据模型建立的需要及数据周期性的特点, 对原始数据进行季节差分等优化处理, 并将建立的模型运用于某水文站流量数据的预测。结果表明, 基于 ARMA 模型对流量数据的预测精度远远高于传统的神经网络模型, 其中基于 AIC 准则定阶的模型比遗传算法定阶的模型精度高 2.96%, 从而为小样本水文数据的预测分析提供一种新的思路。

关键词: 水文数据; ARMA; 定阶准则; AIC; 遗传算法

中图分类号: P333 **文献标识码:** A **文章编号:** 1008-701X(2017)06-0027-04

DOI: 10.13641/j.cnki.33-1162/tv.2017.06.007

The Prediction of Hydrological Data Based on ARMA Model

CHEN Jia-yuan¹, YAN Jie²

(1. Tongxiang Water Conservancy Bureau, Tongxiang 314500, Zhejiang, China;
2. Northwest Electric Power Design Institute Co., Ltd. of China Power Engineering Consulting Group,
Xi'an 710075, Shanxi, China)

Abstract: Because of the difficulty in the prediction of hydrological data for small samples, the article put forward a ARMA model which is used to analyze the periodic hydrological data based on AIC order selection criteria and genetic algorithm. According to the need of the model and the characteristics of data periodicity, the original data were optimized by the seasonal difference, and the model was applied to forecast the discharge data of a hydrological station. Results show that the accuracy of ARMA model is more higher than that of the traditional BP neural networks model. Furthermore, the accuracy of model based on AIC order selection criteria is 2.96% higher than that of the model based on genetic algorithm. Therefore, the article provides a new idea for the analysis of the hydrological data for small samples.

Key words: hydrological data; ARMA; order selection criteria; AIC; genetic algorithm

1 问题的提出

时间序列是按时间顺序排列的、随时间变化且相互关联的数据序列。时间序列分析是根据相邻实测点之间的相互依赖性的特点, 建立数学模型来拟合时间序列, 准确地实现相关信息的有效预测。时间序列分析在经济领域的研究和应用较多。然而, 随着研究的深入, 逐渐扩展到气象、

交通、水利等领域。

水文计算中经常需要长时间序列的水文资料, 但是由于历史、资金等客观原因, 在实际工作中常常无法保证水文资料的完整性, 因此水文数据时间序列的拟合与预测是水文计算中十分关注的重要课题。随着计算机技术以及人工智能的发展, 神经网络及遗传算法等智能算法不断发展, 也越来越广泛地被运用到水文计算当中。如胡军划、唐德善等采用 BP 神经网络编写程序, 根据塔里木河源流叶尔羌河卡群控制水文站历年水文资料, 对径流量进行了预测, 并对模型进行了验证分析^[1]。吴媛、刘国东等针对水文资料的插补和水文资料的延长问题进行了系统研究, 并结合紫坪铺流量资料插补延长资料验证了人工神经网络模型的

收稿日期: 2017-05-25

作者简介: 陈佳袁 (1991-), 男, 工程师, 硕士, 主要从事水利工程管理。

可行性^[2]。余开华结合了神经网络与小波分析分解在函数逼近上的优点, 分析了小波神经网络用于河道流量水位预测的可行性与有效性^[3]。段青青、陈晓楠等人将遗传程序的水位流量关系模型同基于最小二乘法、遗传算法以及蚁群算法的模型进行了详细比较, 表明基于遗传程序设计的水位流量关系模型适应性强, 拟合精度较高^[4]。

但是, 神经网络预测需要大量的数据, 对小样本数据的预测能力相对较差, 模型训练时间不确定, 且会造成局部最优解而非全局最优解; 而遗传算法并不能直接运用于模型的预测, 往往需要与其他模型相结合使用。鉴于上述问题, 本文将时间序列分析法运用于水文数据的预测, 基于水文数据具有的周期性特点, 结合遗传算法将自回归滑动模型 (ARMA) 应用于水文数据的预测, 为水文数据规律性分析提供了一种新的思路。

2 基本原理与方法

时间序列分析方法起源于数学家耶尔, 最初提出的自回归模型 (AR) 主要用于市场变化规律的预测, 在此基础上, 数学家瓦格尔又提出了滑动平均模型 (MA) 和自回归滑动平均模型 (ARMA), 这 3 个模型奠定了时间序列分析法理论基础^[5]。

2.1 ARMA 模型定义

一个系统, 如果时刻 t 时的响应为 X_t , 其值的大小不仅与其之前时刻的值存在关系, 还与之前时刻进入系统的相应扰动存在相关关系, 则这个系统就称为自回归滑动平均系统, 其数学模型为:

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \quad (1)$$

式中: φ_1 为自回归系数; θ_1 为滑动平均系数; $\{\varepsilon_t\}$ 为白噪声序列。

引入后移算子 B , 用 B^k 表示 k 步线性推移算子, 即,

$$B^k X_t = X_{t-k} \quad B^k \varepsilon_t = \varepsilon_{t-k} \quad \text{令}$$

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p \quad (2)$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \quad (3)$$

则式 (1) 简写为:

$$\varphi(B) X_t = \theta(B) \varepsilon_t \quad (4)$$

该模型称为 p 阶自回归 - q 阶滑动平均混合模型, 记为模型 ARMA (p, q), 特殊地, 若时 $p=0$ 为纯滑动平均模型, 记为 MA (q); 若 $q=0$ 时为纯自回归模型, 记为 AR (p), 若 $p=q=0$ 时, 则模型为 $X_t = \varepsilon_t$, 即 $\{\varepsilon_t\}$ 白噪声序列。

2.2 ARMA 模型应用过程

一般来说, 针对动态数据采用 ARMA 模型识别和预测, 其建模过程和相应的方法见图 1。

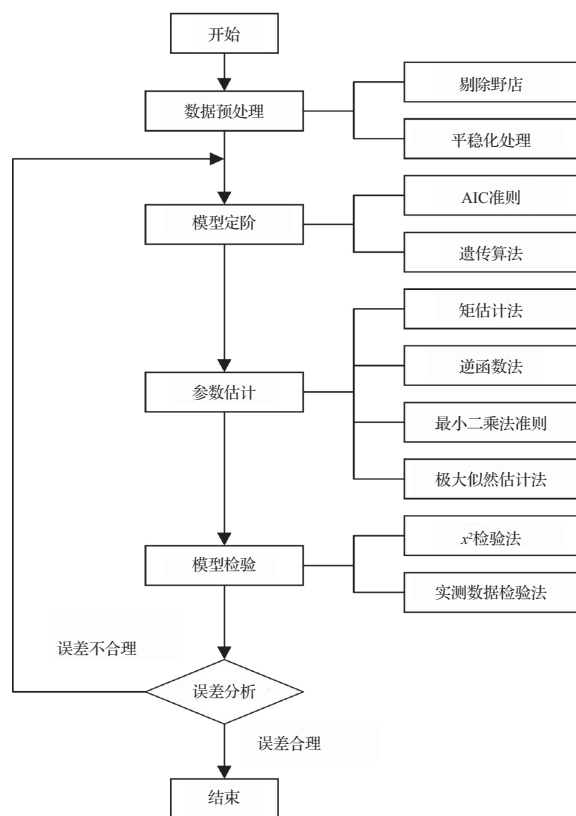


图 1 ARMA 模型建模过程及相应方法图

从图 1 可知, ARMA 模型建模过程主要包括数据预处理、模型定阶、参数估计和模型检验 4 个步骤:

(1) 数据预处理: ARMA 模型构建时需要对数据进行预处理, 主要包括野点剔除和序列平稳化处理。野点剔除的基本思想为: 先从序列中产生一个曲线的平滑估计, 然后把它从数据序列减去, 从而根据设定的值剔除范围之外的点, 根据线性内插补足剔除位置的数据点。序列的平稳化分析可以通过计算样本数据的自相关函数与样本的偏相关函数。如果是截尾或者拖尾的, 则说明数据已经服从 ARMA 模型了。若样本数据不是平稳的, 可以通过差分使数据平稳;

(2) 模型定阶: 确定 ARMA 模型的阶数, 即 p 和 q 的值。阶数的确定可以采用日本统计学家 Akaike 于 1974 年提出的 AIC 准则, 亦可以采用遗传算法定阶^[6];

(3) 参数估计: 确定自回归系数 φ 和滑动平均系数 θ 的值。参数估计的初步估计方法包括矩估计法和逆函数法, 精确估计的方法包括最小二乘法和极大似然估计法。本文采用极大似然估计法, 利用 Matlab 工具箱中的 Garchfit 函数确定相应参数表;

(4) 模型检验: 模型检验的方法包括 x^2 检验法和实测数据检验法。 x^2 检验法的基本思想是: 给定相应的显著性水平, 检验模型的残差序列是否为白噪声序列; 实测数据检验法的基本思想是: 随着实测数据的增加, 将模型预测值与实测数据对比, 计算相应的误差, 从而进一步修改及完善模型。

3 ARMA 模型的应用实例分析

由于水文数据易受气象, 人为等诸多不确定因素的影响, 具有较强的随机性, 因此, 有些水文数据的时间关联性较差, 很难预测。为了更好的分析时间序列分析方法在水文数据中的预测, 本文选取某水文站连续 5 a 的实测月平均流量数据, 该水文站流量变化规律明显, 时间关联性较强, 根据 Tukey 提出的方法剔除野点后得到的流量数据时间序列 $\{X_t\}$ 。剔除野点后水文站实测流量数据见图 2。

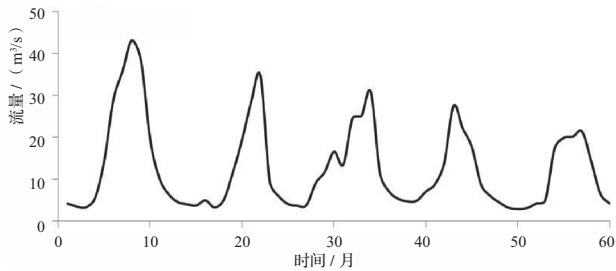


图 2 剔除野点后水文站实测流量数据图

图 2 显示了该水文站连续 5 a 实测的月平均流量变化过程, 由图 2 可知, 该流量数据具有明显的周期性特征且峰值流量呈逐年递减的趋势, 故对流量数据进行季节差分和一阶差分差分运算。季节差分的格式为:

$$\Delta X_t = X_t - X_{t-T} \quad (5)$$

式中: T 为相应时间序列的周期 (月), 当 $T=1$ 时即为一阶差分格式。故通过季节差分和一阶差分后得到的新的时间序列为:

$$\omega_t = \Delta_1 \Delta_T X_t \quad (6)$$

式中: X_t 为实测流量时间序列; Δ_T 为季节差分; 本例中取 $T=12$ 。

经过差分运算后的数据序列已经满足了零均值与平稳性的特点, 差分处理后的实测流量数据见图 3; 且该序列的自相关函数与偏相关函数具有明显的拖尾性, 表示该数据已满足使用 ARMA 模型的要求 (见图 4)。

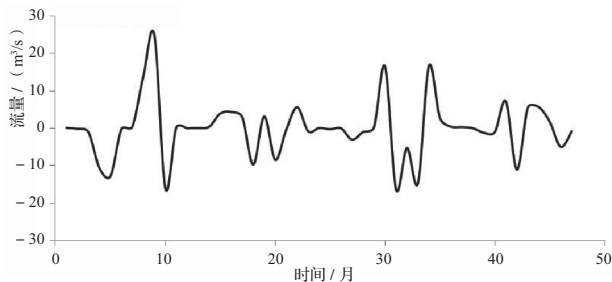


图 3 差分处理后的实测流量数据图

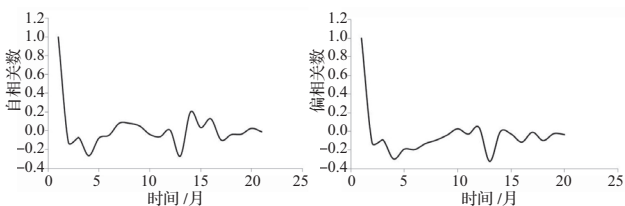


图 4 自相关函数和偏相关函数图

本文建立的 ARMA 模型分别采用 AIC 准则和遗传算法定阶。

(1) 采用 AIC 准则定价, 取 p 的范围为 $0 \sim 14$, q 的范围为 $0 \sim 14$, 经过循环计算后 AIC 的最小值为, 此时对应的 $(p, q) = (1, 1)$, 运用 Garchfit 函数得到 ARMA (1, 1) 模型参数为:

$$\varphi_{AIC}(B) = 1 - 0.5683B \quad (7)$$

$$\theta_{AIC}(B) = 1 - B \quad (8)$$

(2) 采用遗传算法定阶定阶, 将原始流量数据序列分为样本序列和预测序列, 其中前 4 a 的数据为样本数据序列, 第五年的数据为预测数据序列, 取 p 的范围为 $0 \sim 14$, q 的范围为 $0 \sim 14$, 根据样本数据建立模型后, 通过神经网络算法定阶, 运用模型预测第 5 年流量, 并与实测结果比较, 根据误差最小确定最终的优化结果, 最终得到模型的 $(p, q) = (2, 3)$, 经参数估计后得到 ARMA (2, 3) 的参数为:

$$\varphi_{GA}(B) = 1 - 0.6167B + 0.3042B^2 \quad (9)$$

$$\theta_{GA}(B) = 1 - 0.9870B + 0.2705B^2 - 0.2835B^3 \quad (10)$$

通过上述不同的定阶方式和模型参数估计后, 分别建立了基于 AIC 定阶准则的 ARMA (1, 1) 模型和基于遗传算法定阶的 ARMA (2, 3) 模型。

ARMA (1, 1) 模型:

$$\varphi_{AIC}(B) X_t = \theta_{AIC} \varepsilon_t \quad (11)$$

ARMA (2, 4) 模型:

$$\varphi_{GA}(B) X_t = \theta_{GA} \varepsilon_t \quad (12)$$

将上述模型运用于上述水文站第 6 年的流量预测, 模型预测结果见图 5, 并与传统的 BP 神经网络模型预测做相应的比较得到, BP 神经网络建立的模型为:

$$X_t = f(X_{t-1}, X_{t-2}, \dots, X_{t-T}) \quad (13)$$

式中: X_t 为预测值 (m^3/s); T 为周期 (月); 本例中取 $T=12$ 。

不同预测模型的预测结果见图 5。

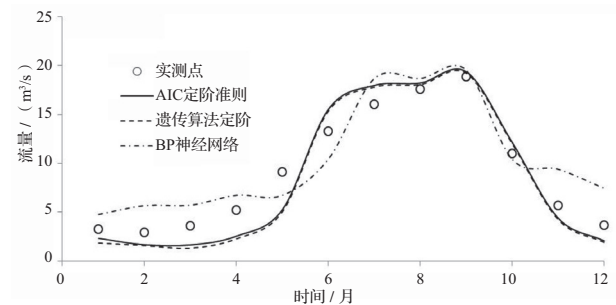


图 5 不同预测模型的预测结果图

并根据下述误差公式计算预测结果的相应误差。得到的误差见表 1。

$$e = \frac{\sum_{i=1}^N \left(\frac{X_i - \hat{X}_i}{X_i} \right)^2}{N} \quad (14)$$

式中: X_i 为流量的实测值 (m^3/s); \hat{X}_i 为流量的预测值 (m^3/s); N 为预测数据的总数。

表 1 模型预测误差对比表

预测方法	ARMA 模型时间序列分析		BP 神经网络
	AIC 定阶准则	遗传算法定阶	
误差 / %	10.93	13.89	26.51

研究表明, 采用 ARMA 模型对流量周期性流量序列的预测具有较高的精度, 由图 5 可知, ARMA 模在小样本时对流量时间序列的预测能很好的反映的了流量丰枯变化, 比传统的 BP 神经网络预测效果更好。由表 1 可知, 基于 AIC 准则定阶的 ARMA (1, 1) 模型预测精度比基于遗传算法定阶的 ARMA (2, 3) 模型精度更高, 其预测相对误差分别为 10.93%, 13.89%, 远远小于传统 BP 神经网络预测的误差 26.51%, 结果表明, 在小样本的周期性水文数据预测中, ARMA 模型的适用性更好。

4 结 论

本文分别基于 AIC 准则和遗传算法建立时间序列分析 ARMA 模型, 并将该模型运用于某水文站流量数据的预测, 并与传统的 BP 神经网络预测模型对比, 得到以下结论:

- (1) 根据流量数据的周期性特定, 可以通过采用特定的季节差分将原始数据处理为零均值平稳序列, 从而满足 ARMA 模型的建模的要求。
- (2) 本文提供了 2 种 ARMA 模型的定阶方法, 基于 AIC 定阶准则和基于遗传算法定阶, 从而建立了 ARMA (1, 1) 模型和 ARMA (2, 3) 模型, 两者定阶建立的模型均

具有较高的预测精度, 其中 ARMA (1, 1) 模型比 ARMA (2, 3) 模型预测误差小 2.96%。
(3) 小样本时间序列的预测中, 通过与传统 BP 神经网络预测模型对比, ARMA 模型预测精度更高, 预测时间更短, 因而更具实用意义。

参考文献:

[1] 胡军华, 唐德善, 胡庆和. 人工神经网络在径流长期预报中的应用 [J]. 人民黄河, 2005 (9): 26 - 27.
[2] 吴媛, 刘国东, 余姝萍, 等. 人工神经网络在水文资料插补延长中的应用 [J]. 贵州地质, 2005 (3): 68 - 71.
[3] 余开华. 小波神经网络模型在河道流量水位预测中的应用 [J]. 水资源与水工程学报, 2013 (2): 204 - 208.
[4] 段春青, 陈晓楠, 韩黎明. 遗传程序在水位流量关系拟合中的应用 [J]. 中国农村水利水电, 2009 (6): 54 - 55.
[5] 梁妍, 夏乐天. 时间序列 ARMA 模型的应用 [J]. 重庆理工大学学报 (自然科学), 2012 (8): 106 - 109.
[6] 崔建国, 赵云龙, 董世良, 等. 基于遗传算法和 ARMA 模型的航空发电机寿命预测 [J]. 航空学报, 2011 (8): 1506 - 1511.

(责任编辑 姚小槐)