文章编号:1000-582X(2012)11-131-05

逐步回归时间序列和 RBF-ANN 在降水预测中的应用

卢文喜[1],杨磊磊[1], 杨忠平[2],辛 欣[1],罗建男[1],初海波[1]
(1.吉林大学 地下水资源与环境教育部重点实验室,长春 130021;
2.重庆大学 土木工程学院,重庆 400044)

摘 要:将逐步回归融入到时间序列预测模型的建立中,摒弃了传统的"考虑所有变量"模式,利用"有进有出"的形式,分清各因子主次关系,仅选用影响显著的变量建立预测方程。径向基函数人工神经网络(RBF-ANN)属于局部逼近网络,准确度高。以桦甸市五道沟站的月降水量和月蒸发量为例,分别用传统、逐步回归时间序列分析和 RBF-ANN 建立降水预测模型,并对比其精度。结果表明:传统、逐步回归时间序列及 RBF-ANN 模型的后验差比值分别为 0.315、0.272、0.284,平均绝对误差分别为 18.37、15.65、13.82 mm,有效系数分别为 0.87、0.94、0.93,精度均满足要求,最后用逐步回归时间序列法预测了未来 5 年的月降水量和月蒸发量。

关键词:时间序列;逐步回归;RBF-ANN;月降水量;预测
中图分类号:TV125 **文献标志码**:A

Application of stepwise regression-time series and RBF-ANN models to precipitation forecasting

LU Wenxi[1], *YANG Leilei*[1], *YANG Zhongping*[2], *XIN Xin*[1], *LUO Jiannan*[1], *CHU Haibo*[1]
(1. Key Laboratory of Groundwater Resources and Environment, Ministry of Education, Jilin University, Changchun, 130021, China;
2. School of Civil Engineering, Chongqing University, Chondqing 400044, China)

Abstract: With integration of stepwise regression into the foundation of time series analysis model, the traditional mode of "take into account all the variables" is abandoned and just significant variables are used to establish the prediction equation in the form of "both enter and exit" mode, with the distinction of each factor's major and minor relationship. The radial basis function artificial neural network (RBF-ANN) belongs to partial approaches network and has high accuracy. Take Huadian County's month precipitation as an example, and compare the accuracy of prediction equations which are established using traditional, stepwise regression time series analysis model and RBF-ANN. The results show that the posterior error ratios of the traditional time series, stepwise regression time series and RBF-ANN models are 0.315, 0.272 and 0.284, the average absolute errors are 18.37 mm, 15.65 mm and 13.82 mm, and the effective coefficients are 0.87, 0.94 and 0.93. At last, we forecast the precipitation and evaporation in future three years with the stepwise regression time series analysis model.

Key words: time series; stepwise regression; RBF-ANN; monthly total precipitation; forecasting

收稿日期:2012-05-03
基金项目:国家自然科学基金资助项目(41072171)
作者简介:卢文喜(1956-),男,吉林大学教授,博士生导师,主要从事地下水数值模拟及水分生态研究,(E-mail) Luwenxi@jlu.edu.cn。

一个地区水资源的丰富程度，取决于降水量的多少[1]。桦甸市多年平均降水量为740.9 mm，多年平均蒸发量为1017.3 mm，水资源相对充沛。但由于地势两翼高，中间低，属于典型的半山区，易发生洪水、泥石流等灾害，因此，对该区降水量进行合理预测，制定相应的防洪抗旱措施，势在必行。

近些年来，降水量的预测得到了极高的重视。韦庆等[1]运用蒙特卡洛法提取已有降水资料的统计特性，作为其内在规律，从而进行预测。李永华等[2]采用BP神经网络预测了汛期的降水量，克服了众多统计方法的限制，揭示了气象体统中非线性的特点。基于水文序列具有时间性、非线性的特点[3]，戴长雷等[4]提出了构建回归分析和时间分析降水预测迭合模型，对资料要求不高，适用性强，其中，因子的选择影响着预测的准确性[5]。

将逐步回归融入时间序列模型，以桦甸市五道沟站实测的月降水量为例，摒弃了传统模式，逐个引入变量，每次在引入新变量之前，对方程中已经存在的变量做显著性检验，剔除不显著变量，保证预测方程中始终都只有显著变量[5-9]。RBF-ANN通过模拟人的大脑神经处理信息的方式，进行信息并行处理，属于局部逼近网络，是多层神经网络中一种常用的网络，避免网络落入局部极小，训练速度较快，准确度较高[10-12]。最终建立了基于逐步回归分析的时间序列预测模型和RBF-ANN模型。

1 时间序列分析模型

1.1 模型原理

时间序列分析就是通过分析已观测的时间序列数据中所蕴含的规律，并利用这些规律来预测未来某一时间段可能达到的水平[3]。通常分别用3种不同的数学方法提取趋势、周期和随机成分，然后将其线性叠加，得到时间序列预测模型[13-14]：

$$P(t) = X(t) + F(t) + R(t), \qquad (1)$$

式中：$X(t)$为趋势项，反映$P(t)$随时间的变化趋势；$F(t)$为周期项，反映$P(t)$的周期性变化；$R(t)$为随机项，反映随机要素对$P(t)$的影响。

用逐步回归多项式拟合法提取趋势分量。把非平稳时间序列趋势成分$X(t)$，($t=1,2,\cdots,n$，n为样本容量)作为因变量，分别将t、t^2等10个因子作为自变量。按对因变量影响的显著程度，把自变量由大到小排序，逐个引入方程，当引入一个新变量时，须对方程中已存在的变量重新检验，剔除不显著变量，保证方程中始终只包含显著变量[5]，最终建立趋势方程。若在选定的显著性水平下，没有因子选入方程，则认为该序列无趋势项[13-14]。

采用谐波分析法，将剩余序列($F_1, F_2, \cdots, F_n$)看成是由不同周期的规则波叠加而成，在分离周期时，逐步分解出一些比较明显的波，然后叠加，作为该时间序列的周期项[13-14]。

消除趋势项和近似周期项后的剩余序列为随机序列项$R(t)$，$R(t)=P(t)-X(t)-F(t)$。考虑到$R(t)$在t时刻的取值与它前1个到p个时间间隔的取值有关，则用自回归模型求解。但$R(t)$并非与所有的R_{t-i}都有显著的关系，因此，用逐步回归法选出对$R(t)$影响较大的R_{t-i}，计算对应系数，得到随机项方程[14-17]。

将趋势、周期、随机分量线性叠加，即可得到降水量的总预测模型。用后验差比值c、小误差频率P进行检验[14-18]，若满足要求，则可用于预测。

1.2 实例应用

选取桦甸市五道沟站实测1990—2010年的月降水资料，用1990—2006年共204个数据计算模型参数，用2007—2010年月降水量检验模型的精度。

1.2.1 趋势项

用编写的Visual Basic 6.0多元逐步回归程序进行计算[19]。结果表明，在显著性水平为0.05时，0个因子选入方程，即该降水序列无趋势项，在研究期内，平均值稳定。

1.2.2 周期项

用编写的Visual Basic 6.0程序计算周期系数[19]，选取最显著的4个波，其结果列于表1。其中，第k个分波对应的周期为n/k，代表该降雨序列的显著周期，有一个周期序号为17，该显著周期为(204/17)12个月，即1年，反映降水量的季节性变化特征；同理可计算另一个显著周期为4～5年，反映降水量的年际波动[13-18]。

表1 傅里叶系数计算结果

显著k	a_0	a_k	b_k
3	66.80	5.43	−6.67
5		−11.94	1.86
17		−71.50	−38.42
34		20.38	30.73

1.2.3 随机项

将剩余序列作为因变量，前一个时间段的降水量作为第1个自变量，以此类推，选用10个预测因子，用多元逐步回归法选取主要因子，结合Visual Basic 6.0程序，计算回归系数，其结果列于表2，相关系数为0.95[6-9,19]。

表 2　回归系数计算结果

Φ_0	Φ_1	Φ_3	Φ_4
25.99	0.31	0.22	−0.17

综上,将 3 个成分线性叠加,得到预测模型。

2　RBF-ANN 模型

2.1　模型原理

RBF-ANN 由输入层、隐含层和输出层组成,如图 1 所示[10-12]。其中,从输入层到隐含层为非线性映射,常采用高斯函数(式 2),从隐含层到输出层为线性映射,RBF-ANN 需要求解的主要参数包括基函数的中心、方差及隐含层到输出层的权值[10-12]。

$$P_i(x) = \exp\left(-\frac{|x-d_i|^2}{2\sigma_i^2}\right), \quad i=1,2,\cdots,n; \tag{2}$$

式中:x 是输入样本;d_i 是第 i 个基函数的中心;σ_i 是高斯函数的方差;n 是感知单元的个数;$\|x-d_i\|$ 是向量 $x-d_i$ 的范数。

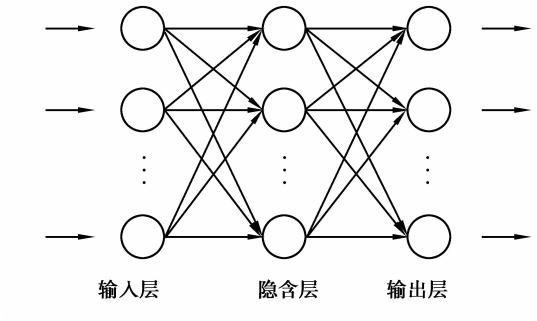

图 1　神经网络结构示意图

2.2　实例应用

将桦甸市五道沟站实测 1990—2010 年的月降水资料,共 252 个数据,作为时间序列 $x=\{x_1,x_2,\cdots,x_{252}\}$。输入层的样本个数为 5,输出层的样本个数为 1,时序划分如表 3 所示,选取前 204 个样本作为训练样本,后 48 个样本作为测试样本。

表 3　时序划分表

样本数	5 个输入	1 个输出
1	x_1、x_2、x_3、x_4、x_5	x_6
2	x_2、x_3、x_4、x_5、x_6x_7	
…	…	…
247	x_{247}、x_{248}、x_{249}、x_{250}、x_{251}	X_{252}

用式(3)对时间序列做归一化处理,用 MATLAB 中的 RBF 工具箱建立网络,先取 SPREAD 的默认值,按 min(mse)找出 GOAL 的最佳区间,使 GOAL 在最佳区间中取值,按 min(mse)搜索 SPREAD 的最佳区间,最终确定最佳训练参数,SPREAD = 1.00,GOAL = 0.003。从而求得 x_{205},x_{206},…,x_{252} 的模拟值[10-12]。

$$y_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \tag{3}$$

式中:$x_{\max}$、$x_{\min}$ 分别为时间序列中的最大、小值;y_i 为 x_i 归一化后的值。

3　模型检验与分析

用 2007—2010 共 4 年的月降水量进行检验,分别计算后验差比值 c、小误差频率 P、平均绝对误差(mean absolute error, MAE)(式 4)及效率系数(efficiency coefficient, CE)(式 5),结果列于表 4 中。由于传统的和逐步回归时间序列分析法大体相似,仅在选用因子时有所差异,所以本文未重复列出计算过程,详见参考文献[13-14]。显然,3 个模型精度均满足要求,都可用于降水预测。

由于降水序列中数值大小相差悬殊,在极值处易造成相对误差过大,因此,出现了几个误差极大点。将逐步回归分析融入到时间序列预测模型的建立中,模型的后验差比值由 0.315 降至 0.272,MAE 由 18.37 降至 15.65 mm,CE 由 0.87 升至 0.94,说明模型经改进后,拟合误差减小,精度得到提高。RBF-ANN 模型和逐步回归时间序列模型的精度相当,时间序列分析模型比 RBF-ANN 模型的 MAE 大,可能是由于降水序列最大值和最小值相差悬殊,在极值处局部预测不准导致整体误差增大。

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}|P_t - \hat{P}_t|, \tag{4}$$

$$CE = 1 - \frac{\sum_{t=1}^{n}(P_t-\hat{P}_t)^2}{\sum_{t=1}^{n}(P_t-\overline{P}_t)^2}, \tag{5}$$

式中:P_t 为实测值;$\hat{P}_t$ 为模型计算值;$\overline{P}_t$ 为实测平均值。

表 4　模型精度比较

方法	c	P	MAE/mm	CE
RBF-ANN	0.284	1	13.82	0.93
时间序列分析模型	0.315	1	18.37	0.87
逐步回归时间序列模型	0.272	1	15.65	0.94

4 模型预测

基于1990—2010年五道沟站的月蒸发资料，用逐步回归时间序列法建立蒸发预测模型，依据建立的降水预测模型和蒸发预测模型，预测了2011—2015共5年的月降水量和月蒸发量(图1)。在预测期，研究区的降水量和蒸发量均以1年为周期，反应了季节性变化。降水和蒸发均没有明显的趋势变化，平均月降水量值为70 mm，平均月蒸发量为86 mm，符合研究区水文气象的一般规律。

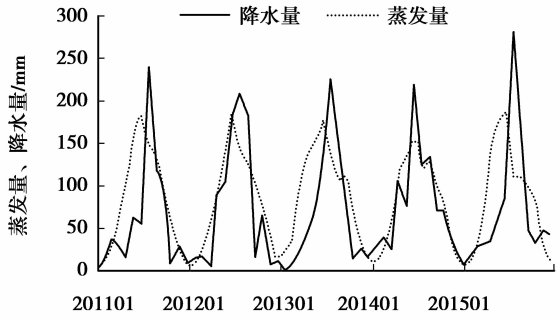

图2 2011—2015年月降水量、月蒸发量预测值

5 结 论

1)以桦甸市五道沟站1990—2010年的月降水量为例，分别用传统、逐步回归时间序列分析和RBF-ANN建立预测模型，经检验，模型精度都合格，但逐步回归时间序列分析和RBF-ANN模型预测效果更佳，能更全面地反映降水的变化特征，进行降水预测。

2)时间序列模型反映了研究区的降水量变化存在2个主要周期：第1个周期长度为4～5年，反映了降水量的年际波动；第2个周期长度为1年，反映了降水量春冬少、夏秋多的特点。研究区雨水过于集中在丰水期，容易形成水土流失、洪涝等自然灾害，水资源得不到充分利用，管理部门应该采取相应的调控措施，进行合理布置。

3)降雨、蒸发的特征变化与当地气候变化、地表特征、地区地形差异等有密切的联系。研究区气象观测站点和降水量测站少，数据缺乏，可能影响到研究结果的精确性。因此进一步的研究需要更多的气象数据来提高研究结果的准确性。

参考文献：

[1] 韦庆，卢文喜，田竹君. 运用蒙特卡罗方法预测年降水量研究[J]. 干旱区资源与环境，2004，18(4)：144-146.
WEI Qing, LU Wenxi, TIAN Zhujun. Application of Monter-Carlo to annual precipitation forecast[J]. Arid Land Resources and Environment, 2004, 18(4): 144-146.

[2] 李永华，刘德，金龙. 基于BP神经网络的汛期降水预测模型研究[J]. 气象科学，2002，22(4)：461-467.
LI Yonghua, LIU De, JIN Long. Study on rainfall prediction model in rain season based on BP Neural Network [J]. Scientia Meteorologica Sinica, 2002, 22(4): 461-467.

[3] 王振龙，胡永宏. 应用时间序列分析[M]. 北京：科学出版社，2007.

[4] 戴长雷，迟宝明，李治军，等. 基于回归分析与时序分析降水预测迭合模型的构建与实现[J]. 河南师范大学学报：自然科学版，2006，34(1)：15-18
DAI Changlei, CHI Baoming, LI Zhijun, et al. Establishment & realization of precipitation forecast model based on regression analysis & time series analysis [J]. Journal of Henan Normal University: Natural Science, 2006, 34(1): 15-18.

[5] 中国科学院数学研究所数理统计组. 回归分析方法[M]. 北京：科学出版社，1974.

[6] 葛朝霞，薛梅，宋颖玲. 多因子逐步回归周期分析在中长期水文预测中的应用[J]. 河海大学学报：自然科学版，2009，37(3)：255-257.
GE Zhaoxia, XUE Mei, SONG Yingling. Application of multi-factor stepwise regression cycle analysis in medium and long-term hydrological forecast [J]. Journal of Hehai University: Natural Sciences, 2009, 37(3): 255-257.

[7] 李辉，练继建，王秀杰. 基于小波分解的日径流逐步回归预测模型[J]. 水利学报，2008，39(12)：1334- 1339.
LI Hui, LIAN Jijian, WANG Xiujie. Stepwise regression model for daily runoff prediction based on wavelet decomposition [J]. Journal of Hydraulic Engineering, 2008, 39(12): 1334-1339.

[8] 索南仁欠. 多元回归分析在水污染评价中的应用[J]. 青海师范大学学报：自然科学版，2004(4)：156-159.
SUONAN Renqian. A study of multivariate statistical analysis on water pollutiou appraises [J]. Journal of Qinghai Normal University: Natural Science Edition, 2004(4): 156-159.

[9] 冯健，张常俊. 均生函数逐步回归模型在水文长期预测中的应用[J]. 东北水利水电，2010(8)：44-46.
FENG Jian, ZHANG Changjun. Application of stepwise regression model based on mean generating function in long-term hydrological forecasting [J]. Water Resources & Hydropower of Northeast China, 2010(8): 44-46.

[10] 艾玲. 时间序列短期预测的方法和技术[D]. 上海：华

东师范大学,2010.

[11] 刘倩然. RBF 人工神经网络在棉花膜下滴灌灌溉预测中的应用[D]. 乌鲁木齐:新疆农业大学,2009.

[12] Gheyas I A, Smith L S. A neural network approach to time series forecasting [C/OL]// Proceedings of the World Congress on Engineering 2009, London, U. K,, July 1-3, 2009 [2012- 01- 10]. http://pdf.aminer.org/000/262/006/forecasting_time_series_combining_machine_learning_and_box_jenkins_time.pdf.

[13] 李平,卢文喜,杨忠平. 频谱分析法在吉林西部地下水动态预测中的应用[J]. 水文地质工程地质,2005,49(4):70-73.
LI Ping, LU Wenxi, YANG Zhongping. Application of spectrum analysis method to the prediction of groundwater regime in west Jilin province [J]. Hydrogeology and Engineering Geology, 2005, 49(4): 70-73.

[14] Yang Z P, Lu W X, Long Y Q, et al. Application and comparison of two prediction models for groundwater levels: a case study in Western Jilin Province China [J]. Journal of Arid Environments, 2009, 73(4/5):487-492.

[15] Gay C, Estrada F, Conde C. Some implications of time series analysis for describing climatologic conditions and for forecasting. An illustrative case: Veracruz, México [J]. Atmósfera, 2007, 20(2):147-170

[16] Yildirim Y E, Turkes M, Tekiner M. Time-series analysis of long-term variations in stream- flow data of some stream-flow stations over the gediz basin and in precipitation of the akhisar station [J]. Palistan Journal of Biological Sciences, 2004, 7(1):17-24.

[17] Zheng Z Q, Fan J S, Liu H P, et al. The analysis and predictions of agricultural drought trend in Guangdong province based on empirical mode decomposition [J]. Journal of Agricultural Science, 2010, 12 (2): 169-174.

[18] Ahn H. Modeling of groundwater heads based on second-order difference time series models. Journal of Hydrology, 2000, 234(1/2):82-94.

[19] 李鸿吉. Visual Basic 6.0 数理统计使用算法[M]. 北京:科学出版社,2003.

(编辑　郑　洁)

(上接第 116 页)

[12] 杨璇,丁百川. 同华煤矿“5·30”事故回顾[J]. 劳动保护,2010, 58(7): 31-33.
YANG Xuan, DING Baichuan. Tonghua coal mine with the “5 · 30” incident review [J]. Labor Protection, 2010, 58(7): 31-33.

[13] 李宗翔. 有源风网模型及其应用计算[J]. 煤炭学报,2010, 35(增刊): 118-122.
LI Zongxiang. Containing the source ventilation network model and its application[J]. Journal of China Soal Society, 2010, 35(Sup): 118-122.

[14] 李宗翔, 王德民, 温永宇. 矿井 3D 风网图及基于 MATLAB 仿真编程与实现[J]. 安全与环境学报, 2010, 10(6): 168-171.
LI Zongxiang, WANG Demin, WEN Yongyu. 3D mine ventilation network graph and MATLAB based simulation program [J]. Journal of Safety and Environment, 2010, 10(6): 168-171.

(编辑　郑　洁)