

一种基于回归分析与时序分析的降水预报模型

唐亚松^{1,2}, 张鑫¹, 蔡焕杰¹, 王健¹

(1. 西北农林科技大学 水利与建筑工程学院, 陕西 杨凌 712100; 2. 武汉大学 水利水电学院, 湖北 武汉 430072)

摘要: 中长期降水预报存在的问题是资料不足和预报精度较低。运用回归分析与时序分析相结合的方法, 将实测降水序列分解成趋势、周期、随机 3 种成分, 并分别构建各分量的子模型; 将 3 个子模型线性叠加, 并对拟合的纯随机序列进行修正, 得到降水预报的第三类模型; 进而给出了模型精度评价方法。实例分析表明, 运用修正后的模型进行降水预报, 可以缩小峰值处的误差, 预报精度比其余两类模型有所提高。该方法可仅根据历史资料进行降水预报, 对资料要求较低, 精度可靠, 是一种实用的方法。

关键词: 降水预报; 回归分析; 时序分析; 模型修正; 精度评价

文献标识码: B

文章编号: 1000—288X(2009)01—0088—04

中图分类号: P457.6

A Precipitation Forecast Model Based on Regression Analysis and Time Series Analysis

TANG Ya-song^{1,2}, ZHANG Xin¹, CAI Huan-jie¹, WANG Jian¹

(1. College of Water Resources and Architectural Engineering, Northwest A & F University, Yangling, Shaanxi 712100, China; 2. College of Water Resources and Hydropower, Wuhan University, Wuhan, Hubei 430072, China)

Abstract: In this paper, a composite model based on regression analysis and time series analysis is constructed to improve forecast accuracy and deal with problems of data deficiencies in mid-to-long-term precipitation forecast. Generally speaking, hydrologic time series may be split into the trend, cyclic, and random components. Under this guideline, firstly, precipitation sequence is split into the three components mentioned above by using regression analysis and time series analysis methods; three submodels are respectively constructed for the three components; the independent random sequence is stabilized to modify traditional model structure; and the three submodels are added by linear superposition to establish a new precipitation forecast model (the 3rd type). Later, the paper gives accuracy assessment indexes. Results show that errors of the modified model are lower in rain spell, i. e., its accuracy is higher than other two models. So, using historical data, the model can work well in precipitation forecast and get a proper accuracy. It is proved to be a practical model for forecast.

Key words: precipitation forecast; regression analysis; time series analysis; model modifying; accuracy assessment

大气降水是一种复杂多变的自然现象, 因而进行中长期降水预报是自然科学领域的一项研究难题。目前降水预报方法有成因分析法、水文统计法、模糊分析法、人工神经网络法等^[1-2]。而科学界对于降水成因及演变规律还处于探索研究阶段, 其它方法应用的问题是资料不足和预报精度较低^[2]。

本研究从水文统计学角度出发, 将回归分析与时序分析方法结合, 并加入纯随机项来进行降水模拟和

预报, 该方法缩小了在峰值处的误差, 可以仅根据历史资料进行预报, 对资料要求低, 预报精度比未考虑纯随机项有一定提高, 预报结果可以给水文工作提供参考, 以及指导农业生产。

1 模型建立

一般而言, 趋势成分、周期成分和随机成分是水文学序列中的 3 种主要成分。因而本研究先提取分段

收稿日期: 2008-05-22

修回日期: 2008-09-29

资助项目: 国家 863 项目 (14110209); 国家重大科技支撑项目 (10712); 西北农林科技大学博士科研启动基金 (01140504); 西北农林科技大学科研专项 (08080230)

作者简介: 唐亚松 (1985—), 男 (汉族), 湖南省临武县人, 硕士研究生, 研究方向为现代水文。E-mail: ty217051@163.com。

通信作者: 张鑫 (1968—), 男 (汉族), 河南省淅川县人, 博士, 副教授, 主要从事水资源可持续开发利用及其生态环境效应方面的研究。E-mail: zhxin@nwsuaf.edu.cn。

趋势,求解趋势分量的数学表达式,得到趋势子模型;然后针对剔除了拟合趋势成分的序列,运用谱分析方法提取并模拟周期分量,得到周期子模型;再对余下的随机成分,建立自回归模型来模拟随机分量,得到随机子模型。最后将 3 个子模型线性叠加得到基本模型,再作适当的修正,得到降水预报模型^[2-5]。

1.1 基本模型的建立

1.1.1 趋势子模型的建立 已有的研究认为^[6],水文序列趋势分析最好只用到年的资料上,且在提取趋势前需做趋势检验,若趋势不明显,则向下作周期分析;若有明显的趋势,才提取趋势项。

据此,对通过趋势检验的序列 $\{X_t, t=1, 2, \dots, n\}$ 采用分段趋势法提取趋势项,计算如下式:

$$C_{q(t)} = \frac{1}{L} \sum_{j=1}^t X_{(q-1) \times L + j} \quad (1)$$

式中: $L=12$ (月); $t=1, 2, \dots, n$, n ——序列长度; $q(t) = \lceil \frac{t-1}{L} \rceil + 1 = 1, 2, \dots, y$, y ——周期的个数, $y = \lceil \frac{n-1}{L} \rceil + 1$; $q(t) = q$, 后面符号意义同此处;符号“ $\lceil \cdot \rceil$ ”表示取整函数。

由此可得到提取的分段趋势序列 $\{C_{q(t)}\}$ 。对该序列,先给出一个一般的多项式函数形式:

$$C_{q(t)} = a_0 + a_1 q(t) + a_2 q^2(t) + a_3 q^3(t) + a_4 q^4(t) + a_5 \frac{1}{q(t)} + a_6 \frac{1}{q^2(t)} + a_7 \sqrt{q(t)} + a_8 \frac{1}{\sqrt{q(t)}} + a_9 \frac{1}{e^{q(t)}} + a_{10} \ln q(t) \quad (2)$$

然后利用逐步回归分析方法^[8]引入对因变量 $C_{q(t)}$ 影响显著的自变量,而不引入对其影响微弱的自变量,以此建立“最优”回归方程,得到趋势子模型,对趋势进行拟合及预报。

1.1.2 周期子模型的建立 确定趋势分量的多项式数学表达式后,从原始数据中剔除拟合趋势 $C_{q(t)}$:

$$P_t = X_t - C_{q(t)} \quad (3)$$

得到序列 $\{P_t, t=1, 2, \dots, n\}$,其中包含了周期成分和随机成分。

用谱分析法^[7]对 $\{P_t\}$ 进行周期分析,找出显著的谐波。提取周期分量 $S_{c(t)}$ 如公式(4):

$$S_{c(t)} = \frac{L}{n} \sum_{k=1}^y (X_{c+(k-1) \times L} - C_k) \quad (4)$$

式中: $c(t) = t - L \times \lceil \frac{t-1}{L} \rceil = 1, 2, \dots, L$; $c(t) = c$; 符号“ $\lceil \cdot \rceil$ ”表示取整函数;其余符号同前。

再按下式求傅立叶系数:

$$\begin{cases} a_0 = \frac{1}{L} \sum_{c=1}^y S_{c(t)} \\ a_i = \frac{2}{L} \sum_{c=1}^y S_{c(t)} \cos \frac{2-i}{L} c(t) \\ b_i = \frac{2}{L} \sum_{c=1}^y S_{c(t)} \sin \frac{2-i}{L} c(t) \end{cases} \quad (5)$$

式中: $i=1, 2, \dots, L/2$,表示谐波的个数。系数的求解可以借用 Excel 完成,求解过程相似,计算方便;也可根据具体情况,选择其中的某些较为显著谐波,纳入模型。

确定傅立叶系数 a_i 和 b_i ,即可确定周期子模型:

$$S_{c(t)} = a_0 + \sum_{i=1}^{L/2} (a_i \cos \frac{2-i}{L} c(t) + b_i \sin \frac{2-i}{L} c(t)) \quad (6)$$

1.1.3 随机子模型的建立 确定周期子模型后,从 P_t 中剔除拟合周期项 $S_{c(t)}$:

$$W_t = P_t - S_{c(t)} \quad (7)$$

得到随机分量序列 $\{W_t\}$,式中 $t=1, 2, \dots, n$; $c(t) = 1, 2, \dots, L$ 。

对序列 $\{W_t\}$ 进行平稳性检验,若序列不平稳,需要进行平稳化处理才能建模;再进行正态性检验,若为正态分布,则采用正态随机数模拟纯随机序列即可,否则需要考虑该序列的偏态特性,采用独立随机项变换法模拟纯随机序列。一般而言,剔除了趋势、周期成分后,余下的随机成分具有平稳性及偏态特性。

考虑到本研究内容是具有相依性的降水量,序列 $\{W_t\}$ 的均值为 0 以及自回归模型 $AR(P)$ 的特点,选用 $AR(P)$ 模型来建立随机子模型,形式如下:

$$W_t = D_t + \sum_{i=1}^p \alpha_i W_{t-i} + \epsilon_t \quad (8)$$

式中: D_t ——相依的平稳随机成分; ϵ_t ——独立的平稳随机成分,即纯随机成分; α_i ——自回归系数; p ——模型的阶数。

模型阶数的确定采用 AIC 准则。具体建模步骤参见文献[7]。

1.1.4 合并子模型 将上述 3 个子模型合并,得到基本模型。一般来说随机分量的自回归模型给出的预测结果为一期望值,所以在预测模型结构中,可省略具有偏态性的纯随机项 ϵ_t ,故基本模型的形式如下:

$$X_t = C_{q(t)} + S_{c(t)} + D_t \quad (9)$$

1.2 模型修正

对于是否考虑纯随机项的问题,分析过程如下。

(1) 第一类模型——忽略纯随机项。首先不考虑 $\{\epsilon_t\}$,计算如(9)式。从模型结构来分析,由于未考虑成分 ϵ_t ,所以相对于实测序列,由公式(9)得到的模拟序列比较平稳,模拟值和实际值在峰值处的差值会比较大。

(2) 第二类模型——加入纯随机项。将具有偏态性的纯随机项 \wedge_t 加入模型,如下:

$$X_t = C_{q(t)} + S_{c(t)} + D_t + \wedge_t \quad (10)$$

从理论上讲, \wedge_t 具有偏态性、随机性,模拟序列中可能会有些不利的结果,比如连续小于 0,连续大于 0 或远大于 0。因此,该模型结构纳入 \wedge_t 后,可能会使模拟序列波动较大。

(3) 第 3 类模型——对纯随机项做修正。针对未考虑 \wedge_t 时,第一类模型模拟序列在峰值处误差大,而考虑 \wedge_t 后模拟序列波动较大的现象,可以对模拟的 \wedge_t 进行修正,剔除绝对值过大的模拟纯随机项 \wedge_t ,使得模拟序列 $\{ \wedge_t \}$ 平稳化。修正方法为:一般未考虑 \wedge_t 时,第一类模型的模拟值大于 0,即 $(C_{q(t)} + S_{c(t)} + D) > 0$ 。在每年少雨期,若加入 \wedge_t 后出现 $[(C_{q(t)} + S_{c(t)} + D_t) + \wedge_t] < 0$,则不引入 \wedge_t ,以此避免出现负值。此时,模型形式如(9)式;若 $(C_{q(t)} + S_{c(t)} + D_t + \wedge_t) / (C_{q(t)} + S_{c(t)} + D) < \alpha$, (α 介于 2~4 间)则引入 \wedge_t ,形式如(10)式。

否则,不引入 \wedge_t ,模型如(9)式。如此即可避免出现模拟序列和实测序列差值大的现象。

加入以上条件后,挑选出符合区域降水特性的模拟序列,要求如下:每年最好只有一个降水量峰值,峰值出现在主要降水时期内;每年的少雨期,降水序列的值较小,无较大的波动。

即可得到利用修正后的模型模拟出的降水序列,该序列比第一类模型有所改进,在主降水期的差值减

$$X_t = 45.82 - 4.22 \ln[q(t)] - (32.35 \cos \frac{c}{6} + 27.72 \sin \frac{c}{6}) - (4.37 \cos \frac{c}{3} - 17.70 \sin \frac{c}{3}) + (9.36 \cos \frac{c}{2} - 5.14 \sin \frac{c}{2}) - (4.95 \cos \frac{2c}{3} - 0.12 \sin \frac{2c}{3}) + (1.26 \cos \frac{5c}{6} + 0.61 \sin \frac{5c}{6}) - 1.50 \cos c + 0.096 W_{t-1} + \wedge_t \quad (11)$$

式中: $q(t) = 1, 2, \dots, y$, y ——周期个数; $c = c(t) = 1, 2, \dots, L$, L ——周期长度; W_{t-1} ——前一个月的随机项; \wedge_t ——模拟的纯随机项, $\wedge_t = \wedge \phi_t$, \wedge ——序列 $\{ \wedge_t \}$ 方差估计, ϕ_t 为服从均值为 0,方差为 1,偏态系数为 C_{ϕ} 的皮尔逊 III 型分布的纯随机变量。

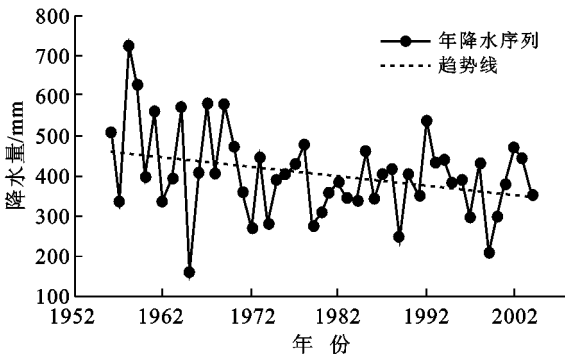


图 1 秃尾河流域高家川站年降水序列图(1956—2004 年)

少,且不会出现第二类模型的情况,预报效果较好。这个模型称为第三类模型。

1.3 模型精度评价方法

采用后验方差检验法及平均绝对百分比误差(MAPE)来检验模型精度^[9-10]。C 值越小则预测模型越好,P 值越大说明误差较小的概率大,模型精度越高。各类精度的等级见表 1。而根据水文情报预报规范,平均绝对百分比误差需在 20% 以内。

表 1 模型预报精度评价查值

预测精度等级	后验方差比值 C	小误差概率 P
一级(特好)	< 0.15	> 0.95
二级(良好)	0.15 ~ 0.25	0.80 ~ 0.95
三级(较好)	0.25 ~ 0.50	0.50 ~ 0.80
四级(合格)	0.50 ~ 0.75	0.25 ~ 0.50
五级(不合格)	> 0.75	< 0.25

2 模型的应用实例

2.1 降水预报模型的建立

据陕西省榆林市秃尾河流域高家川雨量站 1956—2004 年共 49 a 的逐月实测降水资料进行分析,选用前 45 a 的资料建模,后 4 a 资料检验模型预报精度。该站的年降水序列如图 1 所示,对序列作趋势检验,结果表明下降趋势十分显著;对资料进行 3 性审查后,建立趋势、周期、随机分量的子模型,并进行合并。此处给出加入 \wedge_t 的降水预报模型表达式:

以(11)式为基础,对 \wedge_t 作修正,取 2,并挑选出符合该区域降水特性的模拟序列,得到预报效果如图 2。具体修正方法及挑选条件如“模型建立”中所述。

2.2 预报精度评价

为了比较 3 类模型的预报精度,对各模型的预报效果进行了精度评价,各精度指标见表 2;因篇幅有限,省略了前两类模型的预报效果图。

由表 1 的评价标准及表 2 可知,第一类模型预报精度是合格的,模拟值和实际值在峰值处(主要降水期)相差比较大;其余的少雨期,相差不是很大。因此,我们考虑将 \wedge_t 加入到模型中,得到第二类模型。但是该模型在少雨期出现了负值,而且整体上模拟序列的波动性较大,与实测序列拟合效果差,后验方差比指标不合格。因此,需要对 \wedge_t 作出修正。对 \wedge_t 修正,可得到第三类模型。从预报效果图 2 看,第三类

模型纳入纯随机项后,缩小了峰值处的误差,而又没有出现第二类模型所表现出的模拟序列波动无常的情况,少雨期两个序列比较贴近,整体上模拟序列与实测降水序列拟合较好。从各精度指标看,第三类模型的后验方差比值小于前两类模型,说明模拟序列与实测序列比较贴近,预报精度比模型修正前提高了;小误差频率比前两类模型均有提高,达到“较好”等级,说明第三类模型误差较小的概率增大;而平均绝对百分比误差已经小于 10%,误差幅度远小于预报规范要求的 20%,年降水量的预测精度较高。

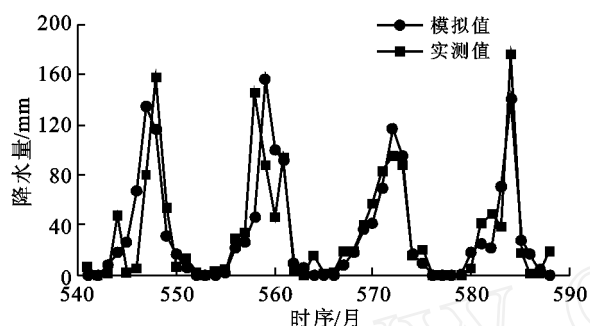


图 2 秃尾河流域高家川站降水预报效果图(第三类模型)

综合上述分析,该模型拟合效果比前两类模型有提高,预报精度介于合格与较好间,效果良好,可以用来进行月降水预报,为水文工作者提供参考,以及指导农业生产。

表 2 3 类模型预报精度对比

模型类型	后验方差比	小误差频率	MAPE/ %
一	0.63	0.65	12.2
二	0.94	0.53	18.6
三	0.61	0.67	6.4

3 结论

(1) 提出了一种降水预报的修正模型。针对第一类模型和第二类模型的缺点,探讨性的提出了一种模型修正方法,将模拟纯随机序列中的异常大和异常小值过滤掉,使之平稳化,提高了模型预报精度。该类模型对资料要求低,而且容易获得,计算简易,模拟精度可靠,是一种实用的预报模型。

(2) 秃尾河流域高家川站的实测降水量整体上有明显的逐年下降趋势;区域降水年内分布不均匀,年际变化较大。将修正后的模型应用于此站,结果表明该模型缩小了与实际序列在峰值处的差距,避免了负值、波动无常的情况;精度指标:后验方差比为 0.61,小误差频率为 0.67,平均绝对百分比误差为 6.4%,这些指标均符合要求,模型预报效果良好,预报结果可以指导当地农业生产。

(3) 利用该模型进行降水预报时,应逐年加入新近收集的水文资料,以使模型吸收新水文因素的变化信息,修正模型及其参数,用新的模型参数进行合理的预报,以获得良好的预报精度。

(4) 所建模型在峰值处拟合精度虽有所提高,但仍存在一定误差,对于未来降水规律的突变还不能较好地反映和模拟出来,这是本方法的缺陷。今后研究中,需要进一步改进模型,纳入实际序列的反常波动信息,提高模型对波动较大的水文序列的识别能力,以改善拟合效果,提高预报精度。

[参 考 文 献]

- [1] 杨旭,栾继虹,冯国章.中长期水文预报研究评述与展望[J].西北农业大学学报,2000,28(6):203-207.
- [2] 戴长雷,迟宝明,李治军.基于回归分析与时序分析降水预报适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.
- [3] 杨忠平,卢文喜,李平.时间序列模型在吉林西部地下水动态变化预测中的应用[J].水利学报,2005,36(12):1475-1479.
- [4] 何书元.应用时间序列分析[M].北京:北京大学出版社,2004.
- [5] Shan Zhao, Wei G W. Jump process for the trend estimation of time series[J]. Computational Statistics & Data Analysis, 2003, 42:219-241.
- [6] 科特戈达(Kodtthoda N T).随机水资源技术[M].金光炎译.北京:农业出版社,1987.
- [7] 丁晶,刘权授.随机水文学[M].北京:中国水利水电出版社,1997.
- [8] 袁志发,周静芋.多元统计分析[M].北京:科学出版社,2003.
- [9] 水利部水利信息中心.中华人民共和国行业标准水文情报预报规范[S].北京:中国水利水电出版社,2001.
- [10] 米鸿燕,蒋兴华.灰色模型 GM(1,1)在建筑物沉降预测中的应用[J].西南林学院学报,2007,27(1):81-83.