

# 基于数理统计方法的降水量预测 模型建立及应用

刘 方, 胡彩虹, 何鹏飞

(郑州大学水利与环境学院, 郑州 450001)

**摘 要:** 降水量是一个随机事件, 但在一个相当长的时间段内又有一定的规律性。由于降水过程存在高度随机性和不确定性, 很难用物理成因等方法来确定某一时段确切的降水量值。在国内具有代表性的权马尔科夫链预测降水量方法的基础上, 结合数理统计的知识, 提出了一种改进的预测降水量思路, 即对原始降水量序列进行 3 a 滑动平均, 并考虑序列间的相关程度, 以减弱原始序列的随机因素, 用新序列进行降水量预测的方法, 并对北京、延安等 5 个站点的降水量序列进行了应用检验。检验结果表明, 除了在极端年份(如崂山站点 2006 年, 偏关站点 2006、2009 年均为枯水年)时预测有较大误差外, 其余年份的预测结果比较令人满意, 总体上合格率达到 80%。由于权马尔科夫链模型建立时在统计学基础上利用了降水量序列的均值和均方差, 预测值是在一定概率条件下趋向于某一状态, 而极端条件发生的概率较小, 因此在预测极端条件时会出现较大误差。

**关键词:** 权马尔科夫链; 降水量预测; 线性回归; 可靠性

**中图分类号:** P457.6

**文献标识码:** A

**文章编号:** 1673-7148(2014)02-0089-05

## 引 言

对于干旱地区来说, 降水是其水资源的主要补给来源, 尤其是在地下水资源开发形势日益严峻的今天。然而降水是一个随机事件, 具有很大的不可预知性, 因此研究历史降水特征以及降水量随时间的变化趋势, 进而总结得到某种内在的联系很有意义。

王文圣等<sup>[1]</sup>介绍的马尔科夫链预测降水的方法中, 仅仅用到一步转移概率, 最终结果为一个各状态概率向量, 概率最大值对应的状态即为预测结果, 没有考虑多年降水序列中前后的相依程度。徐晓宇<sup>[2]</sup>、韩璞璞<sup>[3]</sup>等在马尔科夫链的基础上考虑多年降水量序列中的前后相依关系, 提出权马尔科夫链预测降水量的方法, 但最终结果也是一个状态(3, 平水年), 一个降水量范围。该方法预测的区间范围很大, 没有具体量值, 且对状态的预测有一定的不

可靠性。虽然马尔科夫过程只是对即将发生的过程有影响而对前一个过程无影响, 但是在用权马尔科夫链来预测降水量时, 中间构造的转移概率矩阵是由每一个转移过程在所有历史事件中已经发生的概率得到的, 因此在后来的预测中多少会受到前面某事件发生频率的影响。此外, 廖捷等<sup>[4]</sup>提出的马尔科夫链预测方法, 用不同滞时的转移概率向量与状态均值向量作内积, 再求平均得到最终预测结果, 虽然为具体量值, 但是没有考虑多年降水序列中前后的相依程度。

本文在以上几种预测降水量方法的基础上, 对用马尔科夫链预测降水量的方法进行改进: 在使用权马尔科夫链进行降水量预测时, 首先对降水序列进行 3 a 滑动平均得到新的降水序列, 减弱原始降水序列的随机成分, 再在新的降水序列基础上进行降水预测, 并考虑降水序列内部的相依程度, 由序列内部的相依程度和马尔科夫链各状态转移概率共同

收稿日期: 2013-09-20; 修订日期: 2014-01-08

基金项目: 国家自然科学基金项目(51079131); 国家十二五科技支撑计划(2012BAB); 中国气象局·河南省农业气象保障与应用技术重点开放实验室开放研究基金课题(AMF201304)资助

作者简介: 刘方(1989-), 男, 河南永城人, 本科, 从事水文学及水资源研究, E-mail: liufang1226854@163.com

通讯作者: 胡彩虹(1968-), 女, 山西平遥人, 博士, 教授, 主要从事水文学及水资源方面的教学和科研工作, E-mail: hucaihong@zzu.edu.cn

确定各状态在预测年发生的权重,最后与由各状态平均降水量得到的一系列数值相乘,得到最终预测结果。该方法不仅考虑了降水量序列的内部相依程度,同时考虑预测结果的形式,使预测结果更可信。为了解此模型的可行性与可靠性,利用北京市、安徽省庐江县,以及黄河中游的延安、偏关、岢岚 3 个站点共 5 组降水量数据进行检验。

## 1 基于数理统计方法的降水预测模型建立

数理统计就是对随机变量以及随机变量之间的关系进行定量描述,通过对某些现象发生频率的观察来发现该现象的内在规律性,并做出一定精确程度的判断和预测的数学工具<sup>[5]</sup>。本文建立了基于数理分析方法的一元线性回归、3 a 滑动平均和权马尔科夫链的降水分析及预测模型。

### 1.1 降水量一元线性回归分析

降水的变化趋势可以用一元线性回归方程表示<sup>[6-7]</sup>:

$$y = b_0 + b_1 x \quad (1)$$

式中  $b_0$ 、 $b_1$  为经验回归系数,由最小二乘法求得; $x$  为观测年份; $b_1$  为回归线斜率,表示降水年际变化趋势,一般用  $10 \times b_1$  表示 10 a 的变化趋势<sup>[8-11]</sup>,单位 mm/10a。

### 1.2 滑动平均分析

对序列  $\{x_n\}$  的几个前期值和后期值取平均,得到一个新序列  $\{y_i\}$ ,使原序列光滑化,这就是滑动平均法。数学表达式为

$$y_i = \frac{1}{2k+1} \sum_{i-k}^k x_{i+i} \quad (2)$$

当  $k=1$  时,就是 3 点滑动平均。滑动平均能够清晰、直观地显示出序列的某种趋势并且能减弱随机的因素。

### 1.3 马尔科夫链

马尔科夫链<sup>[12]</sup>是时间和状态都离散的马尔科夫过程。设状态空间  $I = \{1, 2, 3, \dots\}$ ,由转移概率  $P_{ij}$  为元素构成的矩阵为

$$P_k = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{bmatrix} \quad (3)$$

上述矩阵称为  $k$  步转移矩阵。本文只考虑齐次马尔科夫链的情况。

虽然降水量在时间上有一定的规律,但降水量的相依性和独立性,导致其中的随机成分往往被忽略,通常利用自相关性分析<sup>[13-14]</sup>来研究降水序列中

的随机成分。其中各阶自相关系数  $r_k$  计算公式为

$$r_k = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n-k} (x_t - \bar{x})^2} \quad (4)$$

式中  $r_k$  表示第  $k$  阶自相关系数,  $x_t$  表示第  $t$  年的年降水量,  $\bar{x}$  表示逐年降水量序列的多年平均值,  $n$  为年降水量序列的长度。各滞后权重系数  $p_k$  计算公式为

$$p_k = |r_k| / \sum_{k=1}^m |r_k| \quad m \leq 5 \quad (5)$$

式中  $m$  为按预测需要计算到的最大阶数。

### 1.4 预测模型建立

(1) 对原有降水序列进行 3 a 滑动平均得到新的降水序列,从而减弱了原有降水序列的随机因素。

(2) 对新序列进行状态分配,得到各阶各状态的转移矩阵和各状态的平均降水量向量  $a_i$  (第 5 状态为最大值)。

(3) 确定新序列各阶自相关系数,并且对自相关系数进行加权分析,得到各阶自相关系数的权重。

(4) 根据预测使用年份的状态,得到各阶状态转移的权重向量  $p_i$ 。

(5) 各状态均值向量与转移向量作内积 ( $a_i \cdot p_i$ ),得到预测值。

模型预测的合理性用年降水量的绝对误差和相对误差来表示。如果预测结果的相对误差不大于 10% 或绝对误差不大于 50 mm,可以认为预测结果良好,并且优先考虑相对误差的合理性,因为相对误差体现预测值偏离真实值的程度,更加可靠<sup>[15-16]</sup>。

## 2 模型验证及分析

### 2.1 基础资料

数据来源:除了北京市数据来源于百度文库和中新网外,其余数据均来源于相关站点记录数据。北京市数据经过北京市水文局网站以及参考文献[2]对比验证,庐江县数据来源于参考文献[3],延安、偏关、岢岚 3 个站点数据为黄委会水文局整理。

利用北京市、安徽省庐江县,以及黄河中游的延安、偏关、岢岚 3 个站点共 5 组数据,进一步分析其降水变化特征(图 1)。由图 1 可以得出,北京、岢岚年降水量有缓慢减少的趋势,延安、偏关年降水量基本不变,庐江有增加趋势。

### 2.2 北京市预测过程及结果

#### 2.2.1 线性回归和移动平均分析

北京近 63 a (1949—2011 年) 的降水量实测资料一元线性回归模拟和 3 a 移动平均分析见图 2。由图可以看出,北京市年降水量 60 多年来有缓慢减少的

趋势,且每10 a 的减少量大约为45.8 mm。

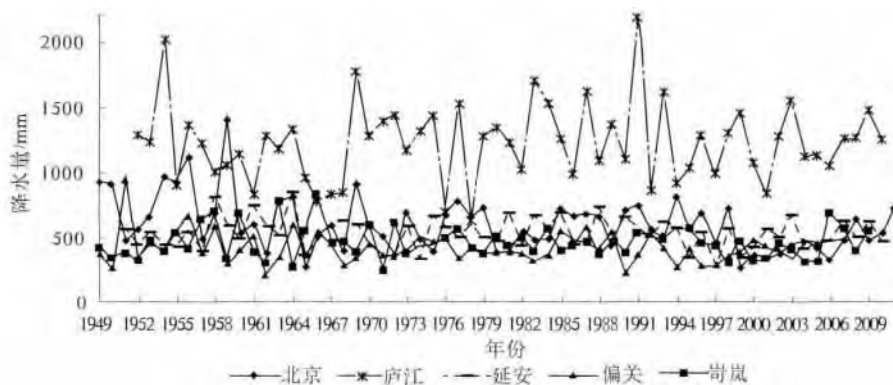


图1 北京、庐江、延安、偏关、岢岚年降水量年际变化

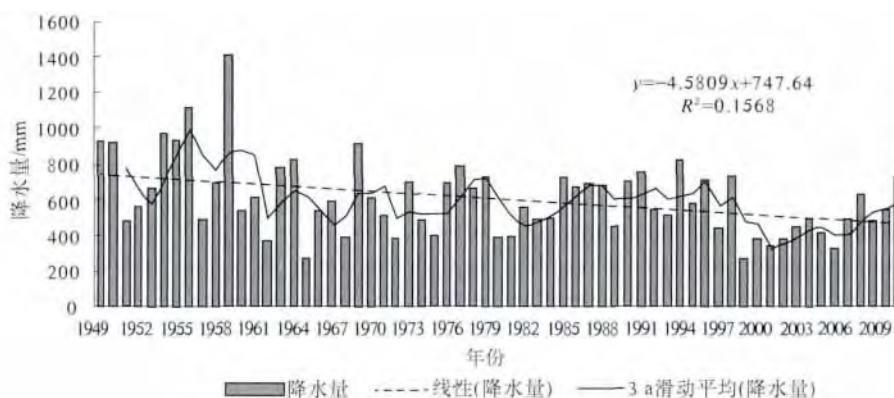


图2 北京市降水量线性回归和3 a 滑动平均图

## 2.2.2 降水预测过程

利用样本均方差对降水量进行分级<sup>[17-19]</sup>。通常按枯水年、偏枯水年、平水年、偏丰水年、丰水年划分为5个等级,分别为 $[0, \bar{x} - a_1s)$ ,  $[\bar{x} - a_1s, \bar{x} - a_2s)$ ,  $[\bar{x} - a_2s, \bar{x} + a_2s)$ ,  $[\bar{x} + a_2s, \bar{x} + a_1s)$ ,  $[\bar{x} + a_1s, +\infty)$ 。 $a_1$ 在 $[1.0, 1.5]$ 中取值, $a_2$ 在 $[0.3, 0.6]$ 取值。其中降水量均值 $\bar{x} = 594.9$ ,标准差 $s = 137.8$ 。这里 $a_1$ 取1.0, $a_2$ 取0.5<sup>[14]</sup>。经过计算,5个区间为: $[0, 457.1)$ 、 $[457.1, 526.0)$ 、 $[526.0, 663.8)$ 、 $[663.8, 732.7)$ 、 $[732.7, +\infty)$ 。将北京市1949—2011年新降水序列按上述区间划分,得到各年降水量状态。然后得到各状态均值向量。

$$\vec{a} = (406 \quad 495.4 \quad 599.4 \quad 698.4 \quad 1003.4)$$

各滞时自相关系数及其权重见表1。根据2007—2011年降水量,对2012年降水量进行预测,过程见

表1 各滞时自相关系数及其权重

$k$	1	2	3	4	5
$r_k$	0.185	0.067	0.184	0.203	0.233
$p_k$	0.3089	0.2215	0.1650	0.1653	0.1393

表2。

由表2得到权向量

$$\vec{p}_i = (0.0926 \quad 0.1489 \quad 0.4475 \quad 0.2198 \quad 0.1242)$$

因此预测2012年降水量为

$$P = \vec{a} \cdot \vec{p}_i = 682.8 \text{ mm}$$

重复以上过程,预测结果见表3。由表3可以看出,预测的5年降水量除2009年外,其余4年的绝对误差量都小于100 mm,且2008年绝对误差量最小,为3.8 mm;除2009年外,其余年份预测相对误差都不大于10%。

表2 降水量预测过程

$n$	$s$	$k$	$p_k$	1	2	3	4	5
61	3	1	0.3089	0	0.1667	0.6250	0.2083	0
60	3	2	0.2215	0.0870	0.1174	0.3913	0.2609	0.1435
59	3	3	0.1650	0.0909	0.2273	0.5000	0.0909	0.1909
58	2	4	0.1653	0.1000	0.1000	0.2000	0.5000	0.2000
57	1	5	0.1393	0.3000	0.1250	0.3750	0	0.2000
权重和				0.0926	0.1489	0.4475	0.2198	0.1242

表 3 北京市降水量预测值与实测值

年份	2008	2009	2010	2011	2012
预测值/mm	642.7	587.2	576.3	652.0	682.8
实测值/mm	628.9	480.6	533.8	721.1	758.6
绝对误差/mm	13.8	106.6	42.5	-69.1	-75.8
相对误差/%	2.2	22.2	8.0	-9.6	-10.0

表 4 延安、偏关降水量预测值与实测值

站点	延安					偏关				
	2006	2007	2008	2009	2010	2006	2007	2008	2009	2010
预测值/mm	522.8	653.2	542.3	622.6	490.8	408.4	502.9	453.5	401.2	514.1
实测值/mm	481.0	620.7	494.2	614.5	469.5	336.4	479.1	467.2	330.0	488.1
绝对误差/mm	41.8	32.5	48.1	8.1	21.3	72.0	23.8	-13.9	71.2	26.0
相对误差/%	8.7	5.2	9.7	1.3	4.5	21.4	5.0	-3.0	21.6	5.3

表 5 岢岚、庐江降水量预测值与实测值

站点	岢岚					庐江县				
	2006	2007	2008	2009	2010	2006	2007	2008	2009	2010
预测值/mm	406.7	633.7	551.4	426.4	554.8	1178.5	1321.3	1311.2	1388.9	1361.4
实测值/mm	301.3	687.3	573.4	393.6	558.8	1052.6	1255.7	1264.7	1480.6	1257.0
绝对误差/mm	105.4	-53.6	-22.0	32.8	-4.0	125.9	65.6	46.5	-91.7	104.4
相对误差/%	35.0	-7.8	-3.8	8.3	0	12.0	5.2	3.7	-6.2	8.3

2.4 预测结果分析

从年降水量预测结果可以看出,预测值与相应年实测值差距不大,相对误差绝对值基本都在 20% 以内,基本满足预测结果要求。而且从表 3、表 4 和表 5 各站点的计算结果可以看出,预测 25 个结果中有 7 个结果(28%)的相对误差绝对值小于 5%,19 个结果(76%)的相对误差绝对值小于 10%;有 21 个结果(84%)的绝对误差绝对值小于 100 mm,其中 14 个结果(56%)绝对误差绝对值小于 50 mm。总体来说,预测结果的满意度接近 80%。因此,本文提出的预测降水量方法是可行的。但是预测的结果并不是都符合要求的,有些结果误差很大,如岢岚站点 2006 年,偏关站点 2006、2009 年属于枯水年,预测结果有较大误差,达不到要求。这是由于权马尔科夫链模型建立时在统计学的基础上利用了降水量序列的均值和均方差,预测值是在一定概率条件下趋向于某一个状态,而极端条件发生的概率较小,在预测极端条件时会出现较大误差。

3 结 论

(1) 本文在现有的权马尔科夫链模型预测降水量的基础上,提出新的预测方法,即先把降水序列进行 3 a 滑动平均,减弱原始序列的随机因素,使降水序列更有规律可循。在得到的新序列上,不仅考虑了新降水量序列内部的相依程度,同时考虑了已发生的各状态降水量均值<sup>[20]</sup>,更最大限度地挖掘了降水

2.3 其余 4 个站点的预测结果

为了增加这种方法的可信度,特选用黄河中游的延安、偏关、岢岚站点,以及安徽省庐江县的年降水量进行预测作为辅助验证,预测结果见表 4、表 5。预测 20 个结果中有 16 个结果(占 80%)的相对误差绝对值小于 10%。综合来说,预测结果比较满意。

量序列的隐含信息。因此,在应用权马尔科夫链预测时,应当先分析降水量的整体特征,寻找隐含的信息,在进行降水量预测时加以使用。

(2) 由于降水的随机性,不可能用公式准确预测出下一时段的降水过程和降水量,如 2012 年虽然能用权马尔科夫链预测出降水量偏多,但不能预测出是哪次降水过程导致 2012 年 7 月北京城区发生严重内涝。用权马尔科夫链预测降水量虽然不能得到准确的过程预测结果<sup>[21]</sup>,但仍有其积极的意义,计算结果可以作为下一阶段开展水资源利用工作的理论参考。

(3) 该方法对降水量有一定的预测能力,但是当遇到枯水年<sup>[22-23]</sup>等特殊年份时用此模型预测会有较大误差,达不到预测要求。

(4) 本文中建立的模型只是从统计学出发进行研究,没有考虑大气动力学的方面,存在一定的局限性。相信随着科技的进步,约束条件的增加使模型越来越完善,降水量预测会越来越准确<sup>[24-26]</sup>。

参考文献

[1] 王文圣,丁晶,金菊良. 随机水文学[M]. 北京:中国水利水电出版社,2008.  
[2] 徐晓宇. 权马尔科夫链在北京市年降水量预测的应用[J]. 科协论坛:下半月,2010(1):111-112.  
[3] 韩璞璞,张生,李畅游,等. 基于权马尔可夫链模型的庐江县降水量预测[J]. 水文,2012,32(3):38-42.  
[4] 廖捷,胡豪然,陈功. 叠加马尔科夫链在年降水量预测中的应用

- [J]. 安徽农业科学 2012 40(9):5532-5533;5604.
- [5]关石菡. 数理统计在数据分析中的应用研究[J]. 林区教学 2011 (6):87-88.
- [6]任露泉. 回归设计及其优化[M]. 北京:科学出版社 2009.
- [7]林元烈. 应用随机过程[M]. 北京:清华大学出版社 2005.
- [8]王春玲, 崔力, 时凤云, 等. 近 54 年濮阳市降水变化特征及突变分析[J]. 中国农业气象, 2010, 31(1):5-10.
- [9]张云霞. 1954-2010 年商丘降水变化趋势及突变分析[J]. 气象与环境科学 2011 34(3):67-73.
- [10]霍正文, 陈文, 凡炳文. 近 54 年定西市降水趋势及突变分析[J]. 水文 2012 32(3):88-92.
- [11]陈敏. 三门峡市近 50 年降水变化趋势及周期分析[J]. 气象与环境科学 2013 36(1):44-49.
- [12]黄振平. 水文统计学[M]. 南京:河海大学出版社 2003.
- [13]王本德. 水文中长期预报模糊数学方法[M]. 大连:大连理工大学出版社 1993.
- [14]袁修开, 吕震宙, 许鑫. 基于马尔科夫链模拟的支持向量机可靠性分析方法[J]. 工程力学 2011 25(2):36-43.
- [15]中华人民共和国水利部. SD138-85 水文情报预报规范[S]. 北京:中国水利水电出版社 2000.
- [16]马占青, 徐明仙, 俞卫阳, 等. 年降水量统计马尔可夫预测模型及其应用[J]. 自然资源学报 2010 25(6):1033-1041.
- [17]张宸, 林启太. 模糊马尔科夫链状预测模型及其工程应用[J]. 武汉理工大学学报 2004 26(11):63-66.
- [18]于延伟, 陈兴伟. 基于 Mann-Kendall 法的径流丰枯变化过程划分[J]. 水资源与水工程学报 2013 24(1):60-63.
- [19]张永民, 肖风劲. 豫西山区降水与气温的波动规律研究[J]. 自然资源学报 2010 25(12):2132-2141.
- [20]闫佰忠, 肖长来, 梁秀娟. 基于数理方法的长岭县降水量分析[J]. 节水灌溉 2012(8):53-56.
- [21]张曙红, 曹建会, 陈锦云. 灰色马尔可夫 SCGM(1,1) 预测模型[J]. 佛山科学技术学院学报:自然科学版, 2004, 22(1):16-19.
- [22]王展, 申双和, 刘荣花. 近 40 a 中国不同两级降水对年降水量变化的影响分析[J]. 气象与环境科学 2011 34(4):7-13.
- [23]于玲玲, 成敏, 廖允成, 等. 马尔科夫链法在分析旱灾规律中的应用[J]. 西北农业学报 2010 19(5):52-56.
- [24]王德智. 水文中长期预报的系统理论和方法研究[D]. 武汉:武汉大学 2003.
- [25]魏凤英. 现代气候统计诊断与预测技术[M]. 北京:气象出版社 1999.
- [26]李森, 夏军, 陈杜明, 等. 北京地区近 300 年降水变化的小波分析[J]. 自然资源学报 2011 26(6):1001-1011.

## Establishment and Application of Precipitation Prediction Model Based on Mathematical Statistics Methods

Liu Fang, Hu Caihong, He Pengfei

(College of Water Conservancy and Environmental Engineering of Zhengzhou University, Zhengzhou 450001, China)

**Abstract:** Precipitation is a random event, but it has certain regularity during a long time period. As the high randomness and uncertain of the precipitation process, it is difficult to use physical genesis method to determine the exact values of a certain period precipitation. Based on domestic representative weighted Markov chain forecast precipitation method, combined with the knowledge of mathematical statistics, an improved precipitation prediction method is proposed. Precipitation prediction can using a new sequence of original precipitation series 3 years moving average, which weakening the random factors of original sequence by considering the correlation between sequences. And precipitation series of 5 stations including Beijing and Yan'an etc. are selected to test. The results show that the station forecast results are rather satisfactory besides in extreme years (such as dry years of Kelan station in 2006, Pianguan station in 2006 and 2009), the prediction of pass rate reaches to 80%. Because of the weighted Markov chain model found by using the mean and mean square error of precipitation sequence based on the statistics, the predicted value is a particular state trend under certain probabilistic condition, and the probability is small for occurring in extreme conditions, so there will be a large error when predicting in extreme conditions.

**Key words:** weighted Markov chain; precipitation prediction; linear regression; reliability