文章编号: 1000-6788(2003)04-0100-06

加权马尔可夫模型在降水丰枯状况预测中的应用

孙才志[1], 张　戈[1], 林学钰[2]
(1. 辽宁师范大学城市与环境学院, 辽宁 大连 116029; 2. 吉林大学资源与环境学院, 吉林 长春 130026)

摘要:　首先基于降水过程存在大量不确定性、不精确性的特点, 应用有序聚类的方法建立降水丰枯状况的分级标准; 然后针对降水量为相依随机变量的特点, 采取以规范化的各阶自相关系数为权重, 用加权的马尔可夫链模型来预测未来降水的丰枯变化状况; 最后以山西省某水文站近 50 年的降水资料为实例对该方法进行了具体的应用, 获得了较为满意的结果.

关键词:　降水; 有序聚类; 权; 马尔可夫链; 预测
中图分类号:　P641.8　　**文献标识码**:　A

Model of Markov Chain with Weights and Its Application in Predicting the Precipitation State

SUN Cai-zhi[1], ZHANG Ge[1], LIN Xue-yu[2]
(1. Urban and Environment Institute, Liaoning Normal University, Dalian 116029, China; 2. Resources and Environment Institute, Jilin University, Changchun 130026, China)

Abstract:　This paper firstly applied sequential cluster method to set up the classification standard of precipitation state based on the fact that there are much uncertainty and imprecise characteristics in the precipitation course; then this paper presented a method which is called Markov chain with weights to predicted the future precipitation state by regarding the standardized self-coefficients as weights based on the special characteristics of precipitation being a dependent stochastic variable; and applied this method to a real hydrological observation station with nearly 50 years precipitation information in Shanxi Province at last, an ideal result was obtained.

Key words:　precipitation; sequential cluster; weight; Markov chain; prediction

1　引言

一个地区降水量的大小, 决定了该地区水资源的丰富程度. 无论是地表水还是地下水, 一般都以大气降水作为最主要的补给来源, 因此在水资源预测、水文预报中经常需要首先对降水量进行预报[1]. 然而, 降水量为一随机变量. 由于气象条件的多样性、变异性和复杂性, 降水过程存在着大量的不确定性、不精确性[2], 从而导致到目前为止还难以通过物理成因来确定出未来某一时段(如年、季、月等)降水量的准确数值. 在实际工作应用中, 仅预测出未来某时段降水量的适当的变化区间即可以完全满足精度要求, 这样一来, 预测的范围扩大了(由点值到区间), 其预测的可靠性也可以相应地提高. 由物理成因的定性分析及大量的降水序列资料的统计分析得知, 降水量为一相依随机变量, 其相依关系的强弱, 通常采用自相关系数作为其定量的测度. 鉴于上述讨论, 我们可以考虑应用有序聚类的方法首先划分出反映降水量丰枯状况的变化区间, 然后以降水量序列规范化后的各阶自相关系数为权, 用加权的马尔可夫链来预测降水量未来的丰枯变化情况.

收稿日期: 2001-12-06
资助项目: 国家重点基础研究项目—973 项目(G1999043606); 辽宁省自然科学基金(001063)
作者简介: 孙才志(1970-), 男, 汉族, 环境科学专业博士后, 副教授, 山东烟台人; 林学钰(1937-), 女, 汉族, 教授, 中国科学院院士, 福建福州人

© 1995-2005 Tsinghua Tongfang Optical Disc Co., Ltd. All rights reserved.

2　有序聚类

有序聚类是对有序样品进行分类的一种方法, 以往通常应用降水量序列的均值与方差的方法来刻画降水量丰枯状况的变化区间. 本文提出应用有序聚类的方法来划分降水量的变化区间, 可以更加充分地考虑降水量序列的数据结构, 使划分的区间更加合理.

有序聚类实现的经典算法是 Fisher 算法[3], 其基本原理为: 设变量 $x_1, \cdots, x_n$ 的某一归类是$\{x_i, \cdots, x_j\}, j \geq i$, 定义其均值向量为

$$\overline{x}_{ij} = \frac{1}{j - i + 1}\sum_{l=i}^{j} x_l \tag{1}$$

将公式

$$D(i, j) = \sum_{l=i}^{j} (x_l - \overline{x}_{ij})'(x_l - \overline{x}_{ij}) \tag{2}$$

定义为$\{x_i, \cdots, x_j\}, j \geq i$ 的直径, 其含义表示该变量段内部各变量之间的差异情况. 其值越小, 表示该段内变量之间差异越小, 或说相互间越接近; 反之, 表示该段内变量之间差异越大, 或说相互间越分散.

设将 n 个有序变量分为 K 类, 某一分法为

$$P(n, K): \{i_1 = 1, \cdots, i_2 - 1\}; \{i_2 = 1, \cdots, i_3 - 1\}; \cdots; \{i_k = 1, \cdots, n\}$$

将公式

$$e[P(n, K)] = \sum_{j=1}^{k} D(i_j, i_{j+1} - 1) \tag{3}$$

定义这一分类的误差函数, 从理论上可以证明, 所谓的最优分法就是使 $e[P(n, K)]$达到最小值时的一种分法. 至于分类数 K 的确定, 可以通过做 $e[P(n, K)]$与 K 关系的曲线图, 曲线拐弯处的 K 值即为最优分类数.

有序聚类的具体原理及详细推导过程可见参考文献[3].

3　权马尔可夫链预测的思想

马尔可夫过程是随机过程的一个分支, 它的最基本特征是"无后效性", 即在已知某一随机过程"现在"的条件下, 其"将来"与"过去"是独立的, 它是一个时间离散、状态离散的时间序列, 其数学表达如下[4-6]:

定义在概率空间(Ω, F, P)上的随机序列$\{X_{(t)}, t \in T\}$, 其中 $T = \{0, 1, 2, \cdots\}$, 状态空间 $I = \{0, 1, 2, \cdots\}$, 称为马尔可夫链, 如果对任意正整数 l, m, k 及任意非负整数 $j_l > \cdots > j_2 > j_1 (m > j_l)$, i_{m+k}, i_m, i_{j_l}, $\cdots$, i_{j_2}, i_{j_1}有

$$\begin{aligned} P\{X_{(m+k)} = i_{m+k} \mid X_{(m)} = i_m, X_{(j_l)} = i_{j_l}, \cdots, X_{(j_2)} = i_{j_2}, X_{(j_1)} = i_{j_1}\} \\ = P\{X_{(m+k)} = i_{m+k} \mid X_{(m)} = i_m\} \end{aligned} \tag{4}$$

成立. 这里, 要求(4)式左端有意义. 即假定

$$P\{X_{(m)} = i_m, X_{(j_l)} = i_{j_l}, \cdots, X_{(j_2)}, X_{(j_1)} = i_{j_1}\} > 0$$

马尔可夫链的性质与特征很多, 在此不一一赘述.

在实际应用中, 一般只考虑齐次马尔可夫链, 即对任意 k, $n \in N^+$, 有

$$P_{ij}(n, k) = P_{ij}(k), \quad i, j = 0, 1, 2, \cdots \tag{5}$$

其中 $P_{ij}(n, k)$表示"于阶段 n 的状态为 i, 经 k 步转移至状态 j 的概率", $P_{ij}(k)$表示"从状态 i 经 k 步转移至状态 j 的概率".

齐次的马尔可夫链$\{X_{(t)}\}$完全由其初始分布$\{P_{(i)}, i = 0, 1, \cdots\}$及其状态转移概率矩阵(状态转移概率 $P_{ij}(i, j = 0, 1, \cdots)$所构成的矩阵)所决定.

由于降水量是一相依的随机变量, 各阶自相关系数刻画了各种滞时的降水量间的相关关系及其强弱, 因而, 可考虑先分别依其前面若干时段的降水量对该时段降水量进行预测, 然后, 按前面各时段与该时段相依关系的强弱加权求和, 即可以达到充分、合理利用信息进行预测的目的, 这就是采用带权马尔可夫链

© 1995-2005 Tsinghua Tongfang Optical Disc Co., Ltd. All rights reserved.

的原因之所在 .

4 权马尔可夫链预测实现的基本步骤

基于上述思路, 权马尔可夫链预测实现的基本步骤为:

1) 计算降水量序列的各阶自相关系数 r_k: $r_k = \sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x}) \Big/ \sum_{t=1}^{x} (x_t - \bar{x})^2$, 式中 x_t 表示第 t 时段的降水量, $\bar{x}$ 表示降水量均值, n 为降水序列长度;

2) 对各阶自相关系数进行归一化, 即, $\tilde{\omega}_k = |r_k| \Big/ \sum_{k=1}^{m} |r_k|$, 并将它们作为各种滞时(步长) 的马尔可夫链的权重(m 为按预测需要计算到的最大阶数);

3) 将降水量序列由小到大排列, 应用有序聚类生成降水量的分级标准;

4) 按 3) 所生成的分级标准, 确定各时段降水量所处的状态;

5) 按 3) 得到的状态序列, 生成不同步长的马尔可夫链的转移概率矩阵;

6) 分别以前面若干时段各自的降水量为初始状态, 结合其相应的状态转移概率矩阵即可预测出该时段降水量的状态概率 $P_i^{(k)}$, i 为状态, $i \in I$, k 为滞时(步长), $k = 1, \cdots, m$;

7) 将同一状态的各预测概率加权和作为降水量处于该状态的预测概率, 即, $P_i = \sum_{i=1}^{m} \omega_k P_i^{(k)}$, 然后将其转化成归一化预测概率 P_i, 在此基础上, 计算各预测时段的状态特征值 $S = \sum_{i=1}^{5} i^{\alpha} P_i$, α 为调整因子, 本次研究取为 4 . 若 $|S - i| < 0.5$, 则该时段降水量所处的预测状态为 i, 待该时段降水量发生后, 将其加入原序列, 重复 1) - 7) 步(再次计算可以省略第三步);

8) 应用马尔可夫链的遍历性定理, 求其极限分布, 进而分析降水量的分布特征 .

5 实例分析

本文以山西省某水文站(国家黄河水利委员会建站) 1952 - 1998 年共 47 年的资料(见表 1) 为例, 进行分析预测, 以说明该方法的具体应用及预测效果 .

表 1 水文站降水序列及状态表

时段	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962	1963
降水量(mm)	261.6	486.4	631.5	259.0	568.0	398.2	479.6	697.6	397.7	640.4	247.1	387.7
状态	1	4	5	1	4	3	4	5	3	5	1	3
时段	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975
降水量(mm)	694.2	211.4	322.6	656.6	325.3	603.8	424.8	383.3	238.8	423.0	237.1	330.7
状态	5	1	2	5	2	5	3	3	1	3	1	2
时段	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987
降水量(mm)	445.9	518.9	492.6	490.3	257.0	400.6	347.5	363.8	411.5	356.2	381.2	317.0
状态	3	4	4	4	1	3	2	2	3	2	3	2
时段	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	
降水量(mm)	473.0	373.7	369.0	348.3	469.4	228.1	338.8	546.1	358.9	412.0	372.3	
状态	4	2	2	2	4	1	2	4	2	3	2	

本次研究先以 1952 - 1995 年 44 年的降水序列预测 1996 年的降水量, 然后将 1996 年实测资料加入到序列中, 再预测 1997 年的降水状态, 最后预测 1998 年的降水状态 .

1) 经计算, 该降水序列各阶自相关系数(对于降水序列, 通常只考虑前 5 阶即可) 分别为: $r_1 = -0.226$, $r_2 = 0.003$, $r_3 = 0.245$, $r_4 = -0.261$, $r_5 = 0.103$;

2) 将各阶自相关系数归一化后作为各种滞时的马尔可夫链的权重, 分别为: $\omega_1 = 0.270$, $\omega_2 = 0.003$, ω_3

© 1995-2005 Tsinghua Tongfang Optical Disc Co., Ltd. All rights reserved.

= 0. 292, ω_4 = 0. 312, ω_5 = 0. 123;

3) 将降水量序列由小到大排列, 经检验(见图 1) 应用有序聚类将降水量分为 5 个区间比较合适, 具体结果见表 2;

4) 确定出各时段降水量的状态(见表 1);

表 2　降水量分级表

状态	级别	数值区间
1	枯水年	$x <$ 289mm
2	偏枯年	289mm $\le x <$ 377mm
3	平水年	377mm $\le x <$ 457mm
4	偏丰年	457mm $\le x <$ 586mm
5	丰水年	$x \ge$ 586mm

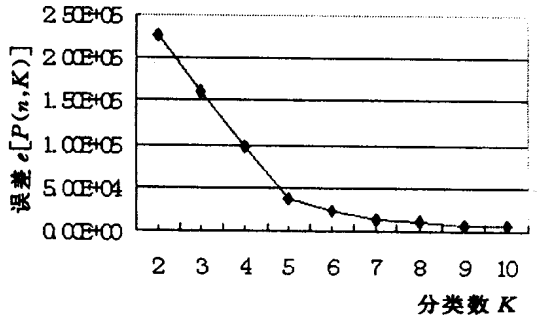

图 1　误差 $e[P(n,K)]$ 与分类数 K 关系曲线图

5) 经计算, 可以得到各种步长的状态转移概率矩阵 P;

6) 根据 1991- 1995 年五年的降水量及其相应的状态转移概率矩阵对 1996 年的降水状态进行预测, 计算结果见表 3;

表 3　1996 年降水量预测表

初始年	滞时(年)	转移概率 状态 / 权重	1	2	3	4	5
1995	1	0. 270	2/8	1/8	1/8	2/8	2/8
1994	2	0. 003	1/10	5/10	2/10	2/10	0/10
1993	3	0. 292	2/7	3/7	0/7	2/7	0/7
1992	4	0. 312	1/7	2/7	3/7	1/7	0/7
1991	5	0. 123	1/9	3/9	2/9	3/9	0/9
P_i 加权求和			0. 210	0. 291	0. 195	0. 237	0. 068
状态特征值			2. 496				

7) 由表 3 可知, S = 2. 496, 这说明 1996 年的降水量状态为 2(偏枯水年), 其值接近于平水年的下限值, 1996 年实测值为 358. 9mm, 与实际情况完全符合. 同理, 以 1992- 1996 年的降水量序列预测 1997 年的降水量状态, 预测结果见表 4

$$P_1 = \begin{vmatrix} 0/8 & 3/8 & 3/8 & 2/8 & 0/8 \\ 0/11 & 3/11 & 3/11 & 3/11 & 2/11 \\ 2/10 & 3/10 & 1/10 & 2/10 & 2/10 \\ 2/8 & 1/8 & 1/8 & 2/8 & 2/8 \\ 3/6 & 1/6 & 3/6 & 0/6 & 0/6 \end{vmatrix} \quad P_2 = \begin{vmatrix} 1/8 & 1/8 & 2/8 & 1/8 & 3/8 \\ 1/10 & 5/10 & 2/10 & 2/10 & 0/10 \\ 3/10 & 2/10 & 2/10 & 2/10 & 1/10 \\ 2/8 & 2/8 & 2/8 & 2/8 & 0/8 \\ 0/6 & 1/6 & 2/6 & 1/6 & 2/6 \end{vmatrix}$$

$$P_3 = \begin{vmatrix} 2/7 & 3/7 & 0/7 & 2/7 & 0/7 \\ 1/10 & 3/10 & 2/10 & 3/10 & 1/10 \\ 1/10 & 3/10 & 5/10 & 1/10 & 0/10 \\ 1/8 & 2/8 & 1/8 & 2/8 & 2/8 \\ 2/6 & 0/6 & 2/6 & 0/6 & 2/6 \end{vmatrix} \quad P_4 = \begin{vmatrix} 0/7 & 1/7 & 2/7 & 2/7 & 2/7 \\ 2/10 & 4/10 & 2/10 & 2/10 & 0/10 \\ 2/10 & 3/10 & 0/10 & 1/10 & 0/10 \\ 1/7 & 2/7 & 1/7 & 2/7 & 2/7 \\ 2/6 & 0/6 & 2/6 & 0/6 & 2/6 \end{vmatrix}$$

© 1995-2005 Tsinghua Tongfang Optical Disc Co., Ltd. All rights reserved.

$$P_5 = \begin{vmatrix} 0/7 & 1/7 & 3/7 & 2/7 & 1/7 \\ 1/9 & 3/9 & 2/9 & 3/9 & 0/9 \\ 2/10 & 4/10 & 3/10 & 1/10 & 0/10 \\ 1/7 & 2/7 & 2/7 & 1/7 & 1/7 \\ 2/6 & 1/6 & 0/6 & 0/6 & 3/6 \end{vmatrix}$$

表 4 1997 年降水量预测表

初始年	滞时(年)	状态 / 转移概率 / 权重	1	2	3	4	5
1996	1	0.270	0/11	3/11	3/11	3/11	2/11
1995	2	0.008	2/8	2/8	2/8	2/8	0/8
1994	3	0.298	1/10	3/10	2/10	3/10	1/10
1993	4	0.305	0/7	1/7	2/7	2/7	2/7
1992	5	0.123	1/7	2/7	2/7	1/7	1/7
加权求和			0.049	0.243	0.257	0.269	0.183
状态特征值			3.323				

由表 4 可知, $S = 3.323$, 这说明 1997 年的降水量状态为 3(平水年), 其值接近于平水年降水量的上限值, 1997 年的实际降水量为 412mm, 与实际情况完全符合. 同理, 以 1993- 1997 年的降水量序列预测 1998 年的降水量状态, 预测结果见表 5.

表 5 1998 年降水量预测表

初始年	滞时(年)	状态 / 转移概率 / 权重	1	2	3	4	5
1997	1	0.270	0/12	4/12	3/12	3/12	2/12
1996	2	0.009	1/11	6/11	2/11	2/11	0/11
1995	3	0.296	1/8	2/8	1/8	2/8	2/8
1994	4	0.303	2/10	4/10	2/10	2/10	0/10
1993	5	0.123	0/7	1/7	3/7	2/7	1/7
P_i 加权求和			0.152	0.298	0.179	0.225	0.145
状态特征值			2.587				

由表 5 可知, $S = 2.587$, 这说明 1998 年的降水量状态为 3(平水年), 但接近于偏枯年的上限值, 而 1998 年实测值为 372.3mm, 属于偏枯年, 虽然降水状态不符合, 但与实际情况基本相符, 这充分说明应用本文的方法进行降水量丰枯状态的预报是可行、有效的.

8) 各种步长的马尔可夫链的特征分析: 由于降水量的马尔可夫链的 5 个状态是相通的, 即 $i \longleftrightarrow j$ ($i, j \in I$, $i \neq j$), 且为非周期的, 其全部状态构成了一个闭集 C, 即该链的状态空间 I, 因而该链是不可约的. 该链为状态空间有穷的不可约马尔可夫链, 故而该链的 5 个状态都是正常返的, 所以该链是遍历的(非周期、不可约、正常返), 根据遍历定理, 可以求出该链的极限分布. 极限分布可以由下面的方程组求出:

$$\begin{cases} p_j = \sum_{k=1}^{n} p_k p_{kj}, \quad j = 1, 2, \cdots, n \\ \sum_{k=1}^{n} p_j = 1 \end{cases} \tag{6}$$

© 1995-2005 Tsinghua Tongfang Optical Disc Co., Ltd. All rights reserved.

本文以相依性最强的步长为 4 的马尔可夫链的特征分析为例, 应用(6)式求出的极限分布为: p_1= 0.152, p_2= 0.303, p_3= 0.244, p_4= 0.190, p_5= 0.111. 根据极限分布可以求出各种状态的再现期, 即 T_i= 1/p_i, 各种状态的再现期分别为: T_1= 6.592(年), T_2= 3.299(年), T_3= 4.096(年), T_4= 5.274(年), T_5= 8.994(年). 由上面的分析可知, 按照本文有序聚类确定的分级标准, 依据现有的资料信息, 在近 50 年的降水过程中, 偏枯水年出现的机会最大, 平均每隔 3.299 年出现一次, 出现的概率为 0.303; 丰水年出现的概率最小, 概率为 0.111, 平均每 8.994 年出现一次; 枯水年出现的概率比丰水年大, 平均每 6.592 年出现一次.

6　结论

在水文、气象科学中, 降水量的预报是一项非常重要的工作, 目前出现了多种预测降水量的方法, 如多元统计、蒙特卡罗模拟、频谱分析等, 但其预测精度都有待提高. 本文的模糊带权马尔可夫链方法, 具有如下的特点

1) 应用有序聚类来确定分级标准, 可以更加充分地考虑降水量序列的数据结构, 从而可以更加有效地刻画降水量序列的内在分布规律, 使划分的降水量区间(分级标准)更加合理.

2) 预测结果为降水量的某一个状态(是一个区间值), 而不是具体数值, 在可以完全满足实际工作需要的前提下, 预测的范围扩大了, 其可靠性会随之有所提高.

3) 由于以各种步长的自相关系数为权, 用各种步长的马尔可夫链加权和来预测降水量状态, 所以较普通的马尔可夫链的预测方法, 它可以更充分、合理地利用信息, 可以成功地将马尔可夫链与相关分析有效的结合起来进行预测.

4) 应用遍历定理, 求计算序列的极限分布, 可以反映出计算序列的许多信息, 从而可以对计算序列进行定性和定量的描述.

5) 如何根据最后计算出的状态概率分布求出降水量的具体值(在某些情况下, 仍需要一个具体数值)仍是一个有待解决的问题. 笔者认为, 由于降水过程的形成是多种模糊因素综合作用的结果, 一方面牵涉到的因素比较多, 另一方面这些因素之间存在着复杂的协同和颉颃作用, 从而形成一个非常复杂的系统, 从这个意义上讲, 将模糊集理论中的隶属度和级别特征值[7](即本文应用的状态特征值)可能是解决这一问题的有效工具.

6) 随着预报对象序列的逐年增加, 资料的代表性也日益增强, 自相关系数、状态转移概率矩阵、权重将会发生某些变化, 这种变化是预报理论模式不断完善的过程, 预报方案不是固定不变的. 因此, 应将每年预报对象的新的实测值加入到资料分析系列, 实现在线调整预报对象的自相关系数、状态转移概率矩阵、权重, 以期进一步提高预报精度[8].

总之, 本文提出的预报降水量的方法将降水成因、统计分析、模糊集分析有机的结合起来, 物理概念清晰, 计算简便, 为提高中长期降水量预报的精度提供了一条值得探索的途径.

参考文献:

[1] 卢文喜. 地下水系统的模拟预测和优化管理[M]. 北京: 科学出版社, 1999. 138-141.

[2] 王文科, 廖健榕. 模糊分析在水文地质学中的应用[M]. 西安: 西安地图出版社, 1997. 1-3.

[3] 胡国定, 张润楚. 多元数据分析方法——纯代数处理[M]. 天津: 南开大学出版社, 1990. 308-312.

[4] 冯耀龙, 韩文秀. 权马尔可夫链在河流丰枯状况预测中的应用[J]. 系统工程理论与实践, 1999, 19(10): 89-93.

[5] Sen Zekai. Critical drought analysis by second-order Markov-order[J]. Journal of Hydrology, 1990, 120(1-4): 183-202.

[6] 周德才, 孙亦鸣. 计算机随机模拟原理、方法及计算程序[M]. 武汉: 华中理工大学出版社, 1998. 201-220.

[7] 陈守煜. 水文水资源系统模糊识别理论[M]. 大连: 大连理工大学出版社, 1992. 4-30.

[8] 陈守煜. 中长期水文预报综合分析理论模式与方法[J]. 水利学报, 1997, (4): 15-21.

© 1995-2005 Tsinghua Tongfang Optical Disc Co., Ltd. All rights reserved.