

加权马尔可夫链在降水状况预测中的应用

夏乐天¹ 朱元甞² 沈永梅¹

(1. 河海大学理学院, 江苏 南京 210098; 2. 河海大学水资源环境学院, 江苏 南京 210098)

摘要 在介绍加权马尔可夫链预测方法的基础上,以郑州市 1951~1994 年降雨量资料为例,采用均值-标准差分级法进行旱涝指标值分级,把这 44 年资料划分成雨涝、偏涝、正常、偏旱、干旱 5 种状态,应用加权马尔可夫链对该地区旱涝状态进行预测和分析,结果表明该方法预测结果与实际情况相吻合。

关键词 加权马尔可夫链;降雨量预测;郑州市

中图分类号:P941 文献标识码:A 文章编号:1006-7647(2006)06-0020-04

Application of weighted Markov chain to prediction of precipitation//XIA Le-tian¹, ZHU Yuan-sheng², SHEN Yong-mei¹
(1. College of Sciences, Hohai University, Nanjing 210098, China; 2. College of Water Resources and Environment, Hohai University, Nanjing 210098, China)

Abstract: Based on the precipitation data for 1951-1994 of Zhengzhou City, the weighted Markov chain is applied to prediction and analysis of states of drought and water-logging in this region, in which the mean and standard deviation of information series are taken as the classification standard of precipitation states, and the past 44 years are classified into five classes according to the precipitation data, i. e. water-logging year, weak water-logging year, normal year, weak drought year, and drought year. It is concluded that the result of prediction by the present method agrees with the reality.

Key words: weighted Markov chain; prediction of precipitation; Zhengzhou City

马尔可夫链是状态和时间均离散的马尔可夫过程,简称“马氏链”。它的最基本特征是“马氏性”,也称“无后效性”,即在系统“现在”的状态已知的条件下,其“将来”的状态与“过去”的状态无关。在现实生活中,很多情况都具有这种属性,例如生物群体的生长与死亡,单位时段内电话交换台来的呼叫次数等。如果具有各种状态的某种事物或某种现象的时间序列可视为马尔可夫链,则根据 n 时刻的状态即可预测 $n+1$ 时刻的状态,这就是应用马尔可夫链模型解决各种预测问题的基本思想。这一基本思想可以应用于天气预报、农作物年景预测、病虫害预测,也可应用于水文、地震和遗传学研究等。本文以郑州市旱涝状态的预测为实例,介绍使用加权马尔可夫链模型进行旱涝状态预测的步骤和方法。

1 马尔可夫链与加权马尔可夫链预测

1.1 马尔可夫链的定义及性质

马尔可夫链的数学表述如下^[1]:定义在概率空间 (Ω, F, P) 上的随机序列 $\{X(t), t \in T\}$,其中参

数集 $T = \{0, 1, 2, \dots\}$, 状态空间 $E = \{0, 1, 2, \dots\}$, 称为马尔可夫链,如果对任意的正整数 l, m, k , 及任意的非负整数 $j_1 > \dots > j_2 > j_1 (m > j_1)$, $i_{m+k}, i_m, i_{j_1}, \dots, i_{j_2}, i_{j_1} \in E$ 有式(1)成立(这里要求式(1)的左端有意义)。

$$P\{X(m+k) = i_{m+k} | X(m) = i_m, X(j_1) = j_1, \dots, X(j_2) = j_2, X(j_1) = j_1\} = P\{X(m+k) = i_{m+k} | X(m) = i_m\} \quad (1)$$

马尔可夫链的性质和有关结论很多,此处不一一叙述。

实际应用中,常记式(1)的右端

$$P\{X(m+k) = i_{m+k} | X(m) = i_m\} = P\{X_{m+k} = j | X_m = i\} = p_{ij}(m, k)$$

$(i, j \in E)$

一般考虑齐次马尔可夫链,即对任意的 $m, k \in T$, 有

$$p_{ij}(m; k) = p_{ij}(k) \quad (i, j \in E) \quad (2)$$

式中 $p_{ij}(m; k)$ 为系统在时刻 m 时处在状态 i 、经 k 步状态转移到达状态 j 的概率; $p_{ij}(k)$ 为系统从状态 i 、经 k 步状态转移到达状态 j 的概率, 此时转移概率与初始时刻无关, k 取 1 时 $p_{ij}(1)$ 记为 p_{ij} 。

齐次马尔可夫链完全由其初始分布 $\{p_0(i_0), i_0 \in E\}$ 及其一步状态转移概率矩阵 $P = (p_{ij})_{i, j \in E}$ 所决定。若已知时刻 n 时的绝对分布 $P(n) = \{p_n(j), j \in E\}$, 则时刻 $n+1$ 时的绝对分布

$$P(n+1) = P(n)P = P(0)P^n \quad (3)$$

1.2 加权马尔可夫链预测

对于一系列相依的随机变量, 用步长为 1 的马尔可夫链模型和初始分布推算出未来时段状态的绝对分布来做预测分析^[2], 即为传统的马尔可夫链预测方法之一, 可称之为基于绝对分布的马尔可夫链预测方法。该法默认所论马尔可夫链满足“齐次性”缺乏依据。事实上, 应用中所论及的随机变量序列, 尽管满足“马氏性”, 但“齐次性”一般都不满足。另外, 该法没有考虑对应各阶(各种步长)马尔可夫链的绝对分布在预测中所起的作用, 因此没有充分利用已知数据资料的信息。利用各阶马尔可夫链求得状态的绝对分布叠加来做预测分析^[3], 也是传统的马尔可夫链预测方法之一, 可称之为叠加马尔可夫链预测方法。该法尽管运用了各阶马尔可夫链状态的绝对分布叠加来预测状态, 但没有考虑各阶马尔可夫链对应的绝对概率在叠加中所起的作用, 即认为各阶马尔可夫链的绝对概率所起的作用是相同的, 这显然不科学。事实上, 一个满足马氏性的相依时间序列, 其各阶自相关性是不一样的。

一系列相依的随机变量, 其各阶自相关系数刻画了各种滞时的状态间相关关系的强弱。因此, 可考虑先分别依其前面若干时段的指标值状态对该时段指标值的状态进行预测, 然后, 按前面各年与该年相依关系的强弱加权求和, 即可达到充分、合理地利用信息进行预测的目的。这就是加权马尔可夫链预测的基本思想。传统的马尔可夫链预测方法和现行的加权马尔可夫链预测方法^[4-5]都没有考虑对指标值序列进行“马氏性”检验, 这是一个欠缺, 本文提出的加权马尔可夫链预测方法考虑到了这个问题, 具体方法步骤如下:

a. 计算指标值序列均值 $\bar{x}$ 和均方差 s , 建立指标值的分级标准(相当于确定马尔可夫链的状态空间), 可根据资料序列的长短及具体问题的要求进行。例如, 可以样本均方差为标准^[2-4]也可以有

序聚类的方法建立分级标准^[5]等, 将指标值分级, 确定马尔可夫链的状态空间 $E = \{1, 2, \dots, m\}$ 。

b. 按步骤 a 所建立的分级标准确定资料序列中各时段指标值所对应的状态。

c. 对步骤 b 所得的结果进行统计, 可得不同滞时(步长)马尔可夫链的转移概率矩阵, 它决定了指标值状态转移过程的概率法则。

d. 进行“马氏性”检验。

e. 计算各阶自相关系数 $r_k(k \in E)$, 计算公式如下:

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

式中: r_k 为第 k 阶(滞时为 k 年的)自相关系数; x_l 为第 l 时段的指标值; $\bar{x}$ 为指标值均值; n 为指标值序列长度。

对各阶自相关系数规范化, 即将

$$w_k = \frac{|r_k|}{\sum_{k=1}^m |r_k|} \quad (5)$$

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)。

f. 分别以前面若干时段的指标值为初始状态, 结合其相应的各阶转移概率矩阵即可预测出该时段指标值的状态概率 $P_i^{(k)}(i \in E)$, k 为滞时(步长), $k = 1, 2, \dots, m$ 。

g. 将同一状态的各预测概率加权和作为指标值处于该状态的预测概率, 即

$$P_i = \sum_{k=1}^m w_k P_i^{(k)} \quad (i \in E) \quad (6)$$

$\max\{P_i, i \in E\}$ 所对应的 i 即为该时段指标值的预测状态。待该时段的指标值确定之后, 将其加入到原序列之中, 再重复步骤 a ~ g, 即可进行下时段指标值状态的预测。

h. 可进一步对该马尔可夫链的特征(遍历性、平稳分布等)进行分析。

2 转移概率的计算、马氏性检验与指标值分级

2.1 马尔可夫链转移概率的计算

为了用马尔可夫链去做预测, 需要根据实测资料估计出马尔可夫链的转移概率。设 $x_1, x_2, \dots, x_n$ 是所讨论的马尔可夫链的 1 个指标值序列, 它包含 m 个状态, 即状态空间 $E = \{1, 2, \dots, m\}$, 用 f_{ij} 表示指标值序列中从状态 i 经过一步转移到达状态 j 的频数($i, j \in E$), 转移的步长可以是 1 个时间单位, 也可以是 2 个、3 个, 甚至 m 个时间单位, 应用上遇

到的马尔可夫链一般不满足“时齐”条件,因此仅需讨论一步转移概率。由 $f_{ij}(i, j \in E)$ 组成的矩阵 $(f_{ij})_{i, j \in E}$ 称为转移频数矩阵。将转移频数矩阵的第 i 行第 j 列元素 f_{ij} 除以各行的总和所得的值称为转移概率,记为 $p_{ij}(i, j \in E)$,即

$$p_{ij} = \frac{f_{ij}}{\sum_{j=1}^m f_{ij}} \tag{7}$$

实际上这是从状态 i 经过一步转移到达状态 j 的转移频率。由频率的稳定性知,当式(7)的分母充分大时,转移频率近似等于转移概率,可用其估计转移概率。因此为方便起见,这里就用转移概率的记号 p_{ij} 来表示转移频率,并且称它为转移概率,于是转移概率矩阵可表示为

$$P = (p_{ij})_{i, j \in E} \tag{8}$$

这里 $p_{ij}(i, j \in E)$ 同式(7)。

2.2 随机变量序列的马氏性检验

检验随机序列是否具有马氏性,是应用马尔可夫链模型分析和解决实际问题的前提。通常离散序列的马尔可夫链可用 χ^2 统计量来检验。

设所讨论的指标值序列包含 m 个可能的状态,用 f_{ij} 表示指标值序列 $x_1, x_2, \dots, x_n$ 中从状态 i 经过一步转移到达状态 j 的频数 $(i, j \in E)$ 。将转移频数矩阵的第 j 列之和除以各行各列的总和所得的值称为边际概率,记为 $p_{\cdot j}$,即

$$p_{\cdot j} = \frac{\sum_{i=1}^m f_{ij}}{\sum_{i=1}^m \sum_{j=1}^m f_{ij}} \tag{9}$$

当 n 充分大时,统计量

$$\chi^2 = 2 \sum_{i=1}^m \sum_{j=1}^m f_{ij} \left| \lg \frac{p_{ij}}{p_{\cdot j}} \right| \tag{10}$$

服从自由度为 $(m-1)^2$ 的 χ^2 分布。其中 p_{ij} 为转移概率,由式(7)定义。

给定显著性水平 α ,查表可得分位点 $\chi^2_{\alpha}((m-1)^2)$ 的值,计算后得统计量 χ^2 的值。若 $\chi^2 > \chi^2_{\alpha}((m-1)^2)$ 则可以认为序列 $\{x_i\}$ 符合马氏性,否则可以认为该序列不可作为马尔可夫链来处理^[6]。

2.3 指标值的样本均值-标准差分级法

对于指标值的分级,传统的方法是应用样本均值与样本标准差来刻画指标值的变化区间,设指标值序列为 $x_1, x_2, \dots, x_n$,样本均值为 $\bar{x}$,那么样本均值标准差为 $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$ 。如果这是一个弱相关序列(相关系数的绝对值小于或等于0.2),则可近似看作是独立同分布的序列。由中心

极限定理知: $P\{\bar{x} - 1.5s \leq x < \bar{x} + 1.5s\} \approx 0.87$; $P\{\bar{x} - s \leq x < \bar{x} + s\} \approx 0.68$ 。于是,可按指标值是否落在 $(-\infty, \bar{x} - 1.0s), (\bar{x} - 1.0s, \bar{x} - 0.5s), (\bar{x} - 0.5s, \bar{x} + 0.5s), (\bar{x} + 0.5s, \bar{x} + 1.0s), (\bar{x} + 1.0s, +\infty)$ 内,将指标值分成5组。应用上常常将指标值的变化区间表示为 $(-\infty, \bar{x} - \alpha_1 s), (\bar{x} - \alpha_1 s, \bar{x} - \alpha_2 s), (\bar{x} - \alpha_2 s, \bar{x} + \alpha_2 s), (\bar{x} + \alpha_2 s, \bar{x} + \alpha_1 s), (\bar{x} + \alpha_1 s, +\infty)$, α_1 可在 $[1.0, 1.5]$ 中取值, α_2 可在 $[0.3, 0.6]$ 中取值。冯耀龙等^[2]就是使用此法对指标值进行分类的。利用这种方法对指标值分类,不考虑物理成因对指标值的影响,仅仅从统计的角度简单地将样本均值作为指标值的中心,在科学性、合理性方面稍显牵强,但操作较为方便,因此应用比较广泛。

3 加权马尔可夫链预测实例

以郑州市1995年旱涝状态的预测为例^[3]介绍预测步骤。郑州市1951~1994年降水量资料如表1所示。

表1 郑州市1951~1994年降水量序列及其状态

年分	降雨量/mm	状态	年分	降雨量/mm	状态
1951	494.5	2	1973	819.6	5
1952	440.3	1	1974	808.2	4
1953	565.2	3	1975	565.9	3
1954	866.5	5	1976	607.7	3
1955	659.3	3	1977	609.1	3
1956	775.6	4	1978	579.1	3
1957	757.7	4	1979	617.7	3
1958	865.0	5	1980	645.4	3
1959	452.6	1	1981	473.4	2
1960	416.6	1	1982	629.4	3
1961	628.7	3	1983	990.6	5
1962	658.0	3	1984	864.4	5
1963	762.8	4	1985	727.0	4
1964	1041.3	5	1986	392.7	1
1965	419.1	1	1987	586.6	3
1966	442.6	1	1988	400.6	1
1967	768.3	4	1989	533.5	2
1968	403.0	1	1990	800.6	4
1969	675.7	3	1991	466.3	2
1970	602.1	3	1992	679.1	3
1971	617.8	3	1993	577.8	3
1972	640.6	3	1994	718.7	4

a. 由表1可计算得年降水量序列均值 $\bar{x} = 637.4$ mm,标准差 $s = 160.47$ mm,取 $\alpha_1 = 1.1, \alpha_2 = 0.5$,按2.3节中的方法可将郑州市年降水量划分为雨涝、偏涝、正常、偏旱和干旱5个状态(参见表2),作为年旱涝状态的标准,据此确定年降水量的分级标准(相当于确定了马尔可夫链的状态空间 $E = \{1, 2, 3, 4, 5\}$)。

b. 按所建立的分级标准确定资料序列中各时

表 2 郑州市 1951 ~ 1994 年降水量分级

状态	级别	分级标准	降雨量区间/mm
1	干旱	$x < \bar{x} - 1.1s$	$x < 460.9$
2	偏旱	$\bar{x} - 1.1s \leq x < \bar{x} - 0.5s$	$460.9 \leq x < 557.2$
3	正常	$\bar{x} - 0.5s \leq x < \bar{x} + 0.5s$	$557.2 \leq x < 717.6$
4	偏涝	$\bar{x} + 0.5s \leq x < \bar{x} + 1.1s$	$717.6 \leq x < 813.9$
5	雨涝	$x \geq \bar{x} + 1.1s$	$x \geq 813.9$

段降水量所对应的状态(见表 1 第 3 列和第 6 列)。

c. 对所得结果进行统计计算,可得不同滞时(步长从 1 到 5)的一步转移概率矩阵,它决定了指标值状态转移过程的概率法则。

d. 对郑州市 1951 ~ 1994 年共 44 年的降水量序列作马氏性检验。由表 1 提供的资料可算得年降水量的频数转移矩阵,再结合步长为 1 的一步转移概率矩阵 $P^{(1)}$,可算得统计量 $\chi^2 = 32.2564$,给定显著性水平 $\alpha = 0.05$,查表可得分位点 $\chi^2_{\alpha}(16) = 26.296$ 。由于 $\chi^2 > \chi^2_{\alpha}((m-1)^2)$,故郑州市年降水量序列满足马氏性。

e. 按式(4)和式(5)可计算得各阶自相关系数 $r_1 = 0.100$, $r_2 = -0.140$, $r_3 = -0.150$, $r_4 = -0.309$, $r_5 = -0.158$,各步长马尔可夫链的权重 $w_1 = 0.117$, $w_2 = 0.163$, $w_3 = 0.175$, $w_4 = 0.361$, $w_5 = 0.184$ 。

f. 依 1994,1993,1992,1991,1990 年的年降水量及其相应的状态转移概率矩阵对 1995 年的年降水量状态进行预测,结果如表 3 所示。

表 3 1995 年年降水量状态预测

初始年	初始状态	滞时/a	权重	状态 1	状态 2	状态 3	状态 4	状态 5	概率来源
1994	4	1	0.117	0.287	0.143	0.143	0.143	0.286	$P^{(1)}$
1993	3	2	0.163	0	0.118	0.471	0.235	0.177	$P^{(2)}$
1992	3	3	0.175	0.063	0.063	0.375	0.250	0.250	$P^{(3)}$
1991	2	4	0.361	0	0	0.667	0.333	0	$P^{(4)}$
1990	4	5	0.184	0.167	0	0.667	0.167	0	$P^{(5)}$

g. 由表 3 可得 $P_1 = 0.075$, $P_2 = 0.047$, $P_3 = 0.522$, $P_4 = 0.250$, $P_5 = 0.106$ 。将同一状态的各预测概率加权求和之后得 $\max\{P_i, i \in E\} = 0.522 = P_3$,可见 1995 年郑州市年降水量的预测状态是 3,即该年的降水量约在 557.2 ~ 717.6 mm 之间,属于正常水文年。实际资料^[2]表明 1995 年郑州市的降水量为 562.5 mm,该结果与预测结果相吻合。

h. 由步骤 e 知,步长为 4 的马尔可夫链的相依性较强,因此可进一步对该马尔可夫链的特征(遍历性、平稳分布等)进行分析。此时的转移概率矩阵是 $P^{(4)}$,由于由 $P^{(4)}$ 所决定的马尔可夫链的 5 个状态是互通的,即对 $\forall i, j \in E, i \leftrightarrow j (i \neq j)$,且为非周期的,其全部状态构成的状态空间是一个闭集,可见此链是

不可约的。因为这是一个有限状态非周期不可约马尔可夫链,因此这是一个正常返链,从而是一个遍历链。根据遍历性定理,可以求出此链的极限分布(此时极限分布即平稳分布)。极限分布求解方程为

$$\sum_{j \in E} \pi_j = 1 \quad \pi_j = \sum_{i \in E} \pi_i p_{ij} \quad (j \in E) \quad (11)$$

由此可解得平稳分布(极限分布)与各状态的重现期如表 4 所示。表 4 中 $\pi_j = 1/\mu_j$ 。

表 4 平稳分布(极限分布)与各状态重现期

状态(j)	π_j	μ_j	状态(j)	π_j	μ_j
1	0.1743	5.74	4	0.2031	4.92
2	0.0816	12.25	5	0.1207	8.29
3	0.4203	2.38			

状态 i 的重现期 $T_i = \mu_i (i \in E)$,对应的概率 $P_i = 1/T_i = \pi_i (i \in E)$ 。

各状态的重现期分别是 $T_1 = \mu_1 = 5.74 \text{ a}$, $T_2 = \mu_2 = 12.25 \text{ a}$, $T_3 = \mu_3 = 2.38 \text{ a}$, $T_4 = \mu_4 = 4.92 \text{ a}$, $T_5 = \mu_5 = 8.29 \text{ a}$ 。由上面的分析可知,按照本文确定的分级标准,根据现有的资料信息,在 40 多年的降水过程中正常年出现的机会最大,平均每隔 2.38 a 出现 1 次,出现的概率为 42.03%;偏旱年出现的机会最小,出现的概率仅为 8.16%,平均每隔 12.25 a 才出现 1 次,偏涝年出现的概率比干旱年和雨涝年的大,平均每隔 4.92 a 就会出现 1 次。

4 结 语

在水文及气象科学中,降水量的预测是一项非常重要的工作,本文所提出的预测方法具有如下特点:

a. 应用样本均值-标准差分级法来确定分级数,再取定 α_1, α_2 适当的值,可以更充分地考虑降水量序列的数据结构,从而更有效地刻画降水量序列内在的分布规律,使划分的降水量区间(分级标准)更加合理。

b. 预测结果为降水量的某一个状态,是一个区间值,而不是一个具体的数值,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,其可靠性也会随之有所提高。

c. 由于以各种步长的自相关系数为权重,用各种步长的马尔可夫链加权来预测降水量状态,所以与普通的马尔可夫链预测方法相比较,本文方法可以更充分、合理地利用信息,成功地将马尔可夫链与相关分析结合起来进行预测。

d. 应用遍历性定理可求出降水量序列的极限分布,它反映出降水量序列的许多信息,从而可进一步对降水量序列作出定性和定量的描述。

(下转第 27 页)

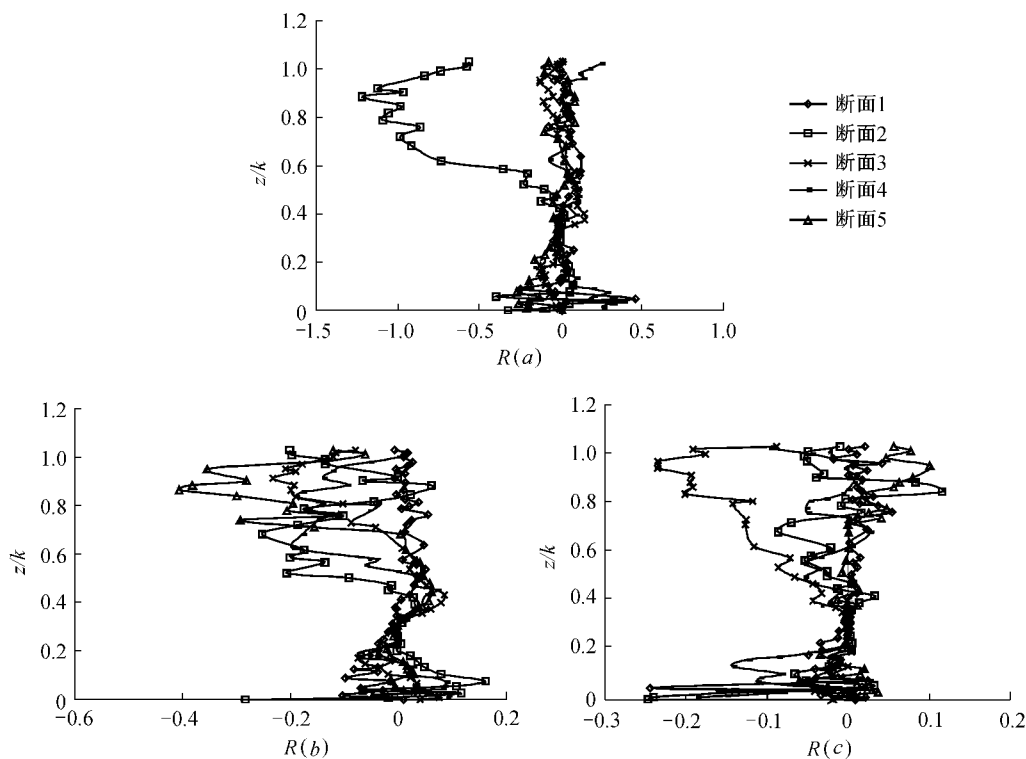


图4 $y = 7.5 \text{ cm}$ 处雷诺应力在不同断面位置的分布

水流分布规律主要与植被本身的结构有关。

b. 在植被区域的顶部有着较强烈的质量和动量交换,纵向、横向、垂向3个方向上的紊动强度相对较大。紊流涡在垂向的统计尺度明显小于在水平方向的尺度,紊动未达到各向同性。另外,不同断面紊动强度的变化规律也不同。

c. 由于植被本身的形状及断面前后植被布置方式的差别,纵向、横向和垂向的雷诺应力在各断面的变化规律不尽相同,流动形态也各异。植被层顶部和

上部水体存在很强的剪切作用,雷诺应力在此处达到最大值。

参考文献:

- [1] COWAN W L. Estimating hydraulic roughness coefficients[J]. Agricultural Engineering, 1956, 37(7): 473-475.
- [2] CHOW V T. Open channel hydraulics[M]. New York: McGraw-Hill, 1959.
- [3] 王付. 河岸植被对水流影响的研究现状[J]. 水资源保护, 2003, 19(6): 50-53.

(收稿日期 2005-05-17 编辑 高建群)

(上接第23页)

e. 如何根据最后计算出的状态概率分布求出所预测的降水量值(在某些情况下需为一个具体的数值),仍是一个有待解决的问题。加权马尔可夫链预测方法必须与其他方法结合起来,才能有效地解决这个问题。

f. 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系列中去,实现动态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。

总之,本文提出的加权马尔可夫链预测方法将序列的自相关分析、多步长一步转移概率分析等有机地结合起来,物理概念清晰,计算简便,为提高中短期降

水量预测的精度提供了一条值得探索的途径。关于马尔可夫链预测方法的优劣,笔者将另文讨论。

参考文献:

- [1] 夏乐天,朱永忠.工程随机过程[M].南京:河海大学出版社,2000:59-81.
- [2] 张汉雄.用马尔可夫链模型预测宁南山区旱情[J].自然灾害学报,1994(3):47-54.
- [3] SEN Ze-kai. Critical drought analysis by second order markov chain[J]. Journal of Hydrology, 1990, 120(1-4): 183-202.
- [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(10):89-93.
- [5] 孙才志,张戈,林学钰.加权马尔可夫链在降水丰枯状况预测中的应用[J].系统工程理论与实践,2003(4):100-105.
- [6] 罗积玉,邢瑛.经济统计分析方法及预测[M].北京:清华大学出版社,1987:347-348.

(收稿日期 2005-08-25 编辑 高建群)