Operations Research and Fuzziology 运筹与模糊学, 2023, 13(3), 1879-1886
Published Online June 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.133187

Hans汉斯

加权马尔可夫链模型对上海地区年降水量的预测

田苏一[1]，孙艺菲[1]，董嘉欣[1]，李宜阳[1]，成常杰[2]

[1]上海工程技术大学数理与统计学院，上海
[2]欧冶云商股份有限公司，上海

收稿日期：2023年4月20日；录用日期：2023年6月13日；发布日期：2023年6月25日

摘　要

采用均值－标准差分级法，以上海地区1970~2020年的年降水量数据为样本降水序列，根据上海地区降水量特点，确定了样本降水序列的分级标准和状态。根据马尔可夫理论和统计学原理，验证了样本降水序列满足马尔可夫性(马氏性)，进而以规范化的各阶自相关系数为权重，建立了适用于该地区降水量的加权马尔可夫链的预测模型。以此模型预测了上海地区2021年和2022年的年降雨量，预测结果比较精确。

关键词

加权马尔科夫链，模糊集理论，预测，降水量，上海地区

Prediction of Annual Precipitation in Shanghai by Weighted Markov Chain Model

Suyi Tian[1], Yifei Sun[1], Jiaxin Dong[1], Yiyang Li[1], Changjie Cheng[2]

[1]School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai
[2]Ouyeel Co., Ltd., Shanghai

Received: Apr. 20th, 2023; accepted: Jun. 13th, 2023; published: Jun. 25th, 2023

Abstract

Using mean-standard deviation classification method, taking annual precipitation data of Shanghai from 1970 to 2020 as sample precipitation series, according to the characteristics of precipitation in Shanghai, the classification standard and status of sample precipitation series were deter-

文章引用: 田苏一, 孙艺菲, 董嘉欣, 李宜阳, 成常杰. 加权马尔可夫链模型对上海地区年降水量的预测[J]. 运筹与模糊学, 2023, 13(3): 1879-1886. DOI: 10.12677/orf.2023.133187

mined. According to Markov theory and statistical principle, it is verified that the sample precipitation series satisfies Markov property. Then, the weighted Markov chain prediction model suitable for the precipitation in this region is established by taking the normalized autocorrelation coefficients of each order as the weight. This model predicts the annual rainfall of Shanghai in 2021 and 2022, and the prediction results are relatively accurate.

Keywords

Weighted Markov Chain, Fuzzy Set Theory, to Predict, Precipitation, Shanghai Area

1. 引言

降水量对农业生产及重大旱涝灾害的产生起着十分重要的作用。但气象条件具有复杂性、多样性和不可预知性，导致降水量在不同地区和不同时间的不均衡性。因此精确的降水量的预报对农业生产和旱涝灾害的控制非常重要。自从 1999 年冯耀龙等[1]首次在我国提出加权马尔可夫链预测的概念以来，国内很多学者利用加权马尔可夫链预测模型成功的对不同地区降雨量进行了预测[2]-[10]。

加权马尔可夫链预测模型是一个步骤清楚、计算方便且客观可靠的方法，值得深入研究和推广。本文从上海气象局官网选取了上海地区 1970~2020 年 51 年的年降水量数据为样本降水序列，采用均值-标准差分级法，根据上海地区降水量特点，建立了适用于上海地区年降雨量的分级标准，并用加权马尔可夫链对上海地区的 2021 和 2022 年的年降雨量状态进行了预测。预测结果所在的状态跟真实的降水量所在的状态一致。再结合模糊集理论的级别特征值，对降水量做了具体预测。结果显示跟真实值相比，平均误差为 4.5%，预测精度较高。

2. 加权马尔可夫模型

马尔可夫过程是一类随机过程，是研究事物的状态和状态转移规律的理论。由俄国数学家 A. A. 马尔可夫于 1907 年提出。它通过不同状态的初始概率和状态之间的转移概率来确定状态的变化趋势，从而达到预测的目的。马尔可夫链具有无后效性特征，这也被后人称为马尔可夫性(“马氏性”)。马尔科夫链预测模型的基本原理。先分别依据前面若干年降水量所对应的状态对某时段的状态做预测，然后按照前面各时段与该时段相关关系的强弱对转移概率进行加权求和。

3. 计算步骤

3.1. 样本分组(均值–均方差法)

上海地区 1970~2020 年的年平均降水量数据 $\{x_1, x_2, \cdots, x_n\}$ 的样本均值和样本均方差分别为 $\bar{x}=1253.7451$，$s=1253.7451$。以样本均值 $\bar{x}$ 为中心，将数据序列分成如下五组(本均值–均方差(标准差)分组法)

$$(-\infty, \bar{x}-\alpha_1 s), [\bar{x}-\alpha_1 s, \bar{x}-\alpha_2 s], [\bar{x}-\alpha_2 s, \bar{x}+\alpha_3 s], [\bar{x}+\alpha_3 s,\ \bar{x}+\alpha_4 s], [\bar{x}+\alpha_4 s, +\infty).$$

根据上海地区降水量特点，取 $\alpha_1=\alpha_4=1.340$, $\alpha_2=\alpha_3=0.185$。依据上面的样本均值–均方差分组法

将 1970~2020 年的年降水量分为 5 个等级(组)，分别为丰涝年、偏丰年、正常年、偏旱年、干旱年，具体分级情况见下表 1。

Table 1. Scale of annual precipitation
表 1. 年降水量分级表

状态	等级	分级标准	降水量区间
1	丰涝年	$x \geq \overline{x} + 1.340s$	$x \geq 1619.4649$
2	偏丰年	$\overline{x} + 0.185s \leq x \leq \overline{x} + 1.340s$	$1304.2363 \leq x < 1619.4649$
3	正常年	$\overline{x} - 0.185s \leq x \leq \overline{x} + 0.185s$	$1203.2539 \leq x < 1304.2363$
4	偏旱年	$\overline{x} - 1.340s \leq x \leq \overline{x} - 0.185s$	$888.0253 \leq x < 1203.2539$
5	干旱年	$x \leq \overline{x} - 1.340s$	$x < 888.0253$

根据表 1 的分级标准，定义将 1970~2020 的降水量所对应状态如表 2 (因表格大小限制，下表中每一行列出 7 年的状态，每行从左到右分别是 7 年对应的状态)。

Table 2. Annual precipitation status table
表 2. 年降水量状态表

年份	状态						
1970~1976	4	4	5	4	4	3	4
1977~1983	2	5	5	2	4	4	2
1984~1990	5	1	4	2	5	2	3
1991~1997	2	4	2	5	3	4	4
1998~2004	3	1	2	1	2	4	4
2005~2011	3	4	3	2	2	2	4
2012~2018	2	4	2	1	2	4	3
2019~2020	1	1					

3.2. “马氏性”检验

采用 χ^2 (卡方分布)来检验该序列是否具有“马氏性”。设 m 为状态数，即 $m=5$。对服从自由度为 $(m-1)^2$ 的 χ^2 分布，给定一个显著性水平 α，查表得 $\chi_\alpha^2\left((m-1)^2\right)$ 的值。如果 $\chi^2 > \chi_\alpha^2\left((m-1)^2\right)$，则拒绝原假设，即认为序列具备“马氏性”。反之，则序列不具备“马氏性”。χ^2 统计量为：

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

其中概率 $p_{ij} = \dfrac{f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij}}$，边际概率 $p_{\bullet j} = \dfrac{\sum_{i=1}^{m} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij}}$ (f_{ij} 表示本年是状态 i 而经过一年后是状态 j 的年份数量)。

如果 $p_{ij}=0$，则 $\ln \dfrac{p_{ij}}{p_{\bullet j}} = -\infty$，且 $f_{ij}=0$，规定 $0\times\infty=0$。

3.3. 计算各阶的相关系数和权重

令 r_k 表示第 k 阶(k 为步长(滞时)， $k \in E = \{1,2,3,4,5\}$)的自相关系数。其公式为：

$$r_k = \frac{\sum_{i=1}^{n-k}(x_i - \overline{x})(x_{i+k} - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \quad k \in E, n = 51.$$

定义权重： $w_k \triangleq \frac{|r_k|}{\sum_{k\in E}|r_k|}$. 故 $\sum_{k\in E} w_k = 1$ 且 $w_k \geq 0$ 。

3.4. 利用各阶转移概率矩阵预测状态

以前面若干时段的状态作为初始状态，利用其相应的各阶转移概率矩阵，即可得该时段的概率 $p_{ij}^{(k)} = 0$，k 为步长。将同一状态的各时段概率加权和作为处于该状态的概率 $P_{(j)} = \sum_{k=1}^{5} w_k p_{ij}^{(k)}$ 。这些概率中最大的 $P_{(i)} = \max_{1\leq j\leq 5}\{p_{(j)}\}$ 所对应的状态即为该时段的预测状态。

3.5. 利用模糊集理论中的级别特征值预测年降水量的值

根据状态概率可得到当年降雨量预测值 X_{n+1}，依据模糊集理论中的级别特征值求法，对各状态定义相应的权重 $W_i = p_i^{\eta} \div \sum_{i=1}^{5} p_i^{\eta}$, $1 \leq i \leq 5$，η 为最大概率的作用系数，通常取为 2 或 4。级别特征值 $H = \sum_{i=1}^{5} i \cdot W_i$ 。如果年降雨量预测状态为 i，则年降雨量预测值为 $X_{n+1} = \frac{HD_i}{i}$，期中 D_i 为状态 i 区间值的下限。

上面公式在应用时结果并不十分理想，因此本文在大量计算的基础上对公式进行了分类改进[9]。

$$\begin{cases} X_{n+1} = \dfrac{HD_i}{i-1.2}(i=4) \\ X_{n+1} = \dfrac{HD_i}{i+0.2}(i=3, H \geq 3.6) \\ X_{n+1} = \dfrac{HD_i}{i-0.2}(i=3, H < 3.6) \\ X_{n+1} = \dfrac{HD_i}{i+1}(i=2, H > 2.8) \\ X_{n+1} = \dfrac{HD_i}{i+0.3}(i=2, H \leq 2.8) \\ X_{n+1} = \dfrac{HD_i}{i+0.8}(i=1, \ H > 2) \end{cases}$$

4. 计算各步长的转移概率矩阵

4.1. 步长为 1 年的转移概率矩阵

取步长为 1 年，对 1970~2020 这 51 年进行统计总结状态变化次数如表 3 (如果本次是 1970 年的状态，下次状态就是 1971 年的状态。表 3 中状态 $i, 1 \leq i \leq 5$ 所在的行与 $j, 1 \leq j \leq 5$ 所在的列交叉位置的数字记为 f_{ij}，表示本年是状态 i 而下一年是状态 j 的年份数量。如 $f_{11} = 1, f_{25} = 4, \cdots$)。

Table 3. Statistical table of precipitation state of step size 1
表 3. 步长为 1 年的降水量状态统计表

本次 \ 下次	状态 1	状态 2	状态 3	状态 4	状态 5
状态 1	1	3	0	1	0
状态 2	2	2	1	6	4
状态 3	2	2	0	3	0
状态 4	0	6	5	5	1
状态 5	1	2	1	1	1

由表 3 可得步长为 1 年的转移概率矩阵

$$P_1=\begin{bmatrix} 0.2 & 0.6 & 0 & 0.2 & 0 \\ 0.1333 & 0.1333 & 0.0667 & 0.4 & 0.2667 \\ 0.2857 & 0.2857 & 0 & 0.4286 & 0 \\ 0 & 0.3529 & 0.2941 & 0.2941 & 0.0589 \\ 0.1667 & 0.3333 & 0.1667 & 0.1667 & 0.1667 \end{bmatrix}。$$

根据表 3 中的数据计算可得各种状态的边际概率(见表 4)

Table 4. The marginal probability of each state
表 4. 各种状态的边际概率

状态	1	2	3	4	5
边际概率	0.10	0.3	0.14	0.34	0.12

4.2. “马氏性”检验

由表 3 和表 4 及步长为 1 年的转移概率矩阵，计算得卡方分布 $\chi^2=27.9988$。给定显著性水平 $\alpha=0.05$，自由度 $(m-1)^2=(5-1)^2=16$。查表可知 $\chi^2_{0.05}(16)=26.296$，$\chi^2>\chi^2_{0.05}(16)$，拒绝原假设，满足“马氏性”。

4.3. 各阶的自相关系数和权重

由 3.3 中公式得：(表 5)

Table 5. Autocorrelation coefficient and weight of each order
表 5. 各阶的自相关系数和权重

阶数	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
r_k	0.0036	0.0944	0.0638	0.0798	0.0506
w_k	0.0123	0.3230	0.2182	0.2732	0.1732

4.4. 步长为 2~5 年的转移概率矩阵

类似于步长为 1 年的情形，可得步长为 2~5 年的转移概率矩阵分别为：

$$P_2=\begin{bmatrix}0.25 & 0.25 & 0 & 0.5 & 0\\ 0.0667 & 0.5333 & 0.1333 & 0.2 & 0.0667\\ 0.1429 & 0.4286 & 0.1429 & 0.2857 & 0\\ 0.1765 & 0.1176 & 0.1765 & 0.2353 & 0.2941\\ 0 & 0.1667 & 0.1667 & 0.6667 & 0\end{bmatrix}\quad P_3=\begin{bmatrix}0 & 0.25 & 0.25 & 0.25 & 0.25\\ 0.1333 & 0.2 & 0.1333 & 0.4667 & 0.0667\\ 0.1667 & 0.5 & 0.1667 & 0 & 0.1667\\ 0.1765 & 0.3529 & 0.1176 & 0.2353 & 0.1176\\ 0 & 0.3333 & 0.1667 & 0.5 & 0\end{bmatrix}$$

$$P_4=\begin{bmatrix}0.25 & 0.25 & 0.25 & 0.25 & 0\\ 0.0667 & 0.4 & 0.1333 & 0.3333 & 0.0667\\ 0.1667 & 0.3333 & 0 & 0.1667 & 0.3333\\ 0.1875 & 0.3125 & 0.1875 & 0.25 & 0.0625\\ 0 & 0.1667 & 0.1667 & 0.5 & 0.1667\end{bmatrix}\quad P_5=\begin{bmatrix}0.25 & 0 & 0.25 & 0.5 & 0\\ 0.2143 & 0.0714 & 0.2143 & 0.3571 & 0.1429\\ 0 & 0.0667 & 0.1667 & 0.1667 & 0\\ 0.0625 & 0.375 & 0.125 & 0.3125 & 0.125\\ 0.1667 & 0.6667 & 0 & 0 & 0.1667\end{bmatrix}$$

5. 结果预测

5.1. 2021 年降水量的预测结果

用 2016~2020 年的年降水量状态，对应的状态转移概率矩阵及表 5 对 2021 年的年降水量状态进行预测，结果见表 6：

Table 6. Using 2016~2020 data to predict precipitation status in 2021
表 6. 用 2016~2020 年的数据预测 2021 年的降水量状态

年份	状态	权重状态	1	2	3	4	5
2020	1	0.0123	0.2	0.6	0	0.2	0
2019	1	0.3230	0.25	0.25	0	0.5	0
2018	3	0.2182	0.1667	0.5	0.1667	0	0.1667
2017	4	0.2732	0.1875	0.3125	0.1875	0.25	0.0625
2016	2	0.1732	0.2143	0.0714	0.2143	0.3571	0.1429
	加权和		0.2079	0.2950	0.1247	0.2941	0.0782

由表 6 得 0.2950 最大，其对应的状态为 2。故预测 2021 年降水量所对应的状态为 2。而根据得到的数据中可以得知 2021 年的降水量为 1478.5，对应的为状态 2，预测成功。根据模糊集理论计算可得 $H = 2.69$，年降水量 1523.40。与实际降水量相对误差为 3.0%。预测效果较好。

5.2. 2022 年降水量的预测结果

同样的方法利用 2017~2021 年的年降水量数据，可以预测 2022 年降水量。

由表 7 可知，0.3130 最大，因此最大加权预测 2022 年降水量所对应的状态为 4。而根据得到的数据中可以得知 2022 年的降水量为 1079.4，对应的为状态 4，预测成功。根据模糊集理论计算可得 $H = 3.20$，年降水量 1014.97。与实际降水量相对误差为 5.9%。预测效果较好。

Table 7. Using 2017~2021 data to predict precipitation status in 2022
表 7. 用 2017~2021 年的数据预测 2022 年的降水量状态

年份	状态	权重状态	1	2	3	4	5
2021	2	0.0796	0.0714	0.1429	0.0714	0.4286	0.2857
2020	1	0.3041	0.2	0.4	0	0.4	0

Continued

2019	1	0.2106	0	0.25	0.25	0.25	0.25
2018	3	0.2372	0.1667	0.1667	0	0.3333	0.3333
2017	4	0.1685	0.0588	0.3529	0.1176	0.3529	0.1176
	加权和		0.1160	0.2918	0.0781	0.3130	0.1743

5.3. 遍历性与平稳分布

以相依性最强的步长为 2 的马尔可夫链进行分析，由于降雨量的 5 个状态是互通的，没有周期性。因此这是一个不可约的正常返的马尔可夫链。此链具有遍历性，其极限分布即为为平稳分布。令状态为 i，j 的平稳分布分别为 π_i, π_j，各状态重现的周期为 T_i，则有方程组：

$$\begin{cases} \pi_j = \sum_{i=1}^{5} \pi_i p_{ij} \\ \sum_{i=1}^{5} \pi_j = 1, \pi_j \geq 0 \end{cases}。$$

设各状态重现的周期为 T_i $\left(T_i = 1/\pi_i\right)$，根据步长为 2 的状态转移矩阵，可得 π_j 与 T_j，见表 8：

Table 8. Limit distribution and recurrence period of each state
表 8. 极限分布与各状态重现周期

状态	1	2	3	4	5
π_j	0.0906	0.3254	0.1996	0.2803	0.1041
T_j	11.0375	3.0371	5.0100	3.5676	9.6061

由表 8 可知，按照本文的分级标准，在 1970~2020 年共 51 年的降水过程中，偏丰年出现的概率最大，平均每隔 3.0371 年出现一次；丰涝年出现的概率最小，平均每隔 11.0375 年才出现一次。

6. 模型拓展训练

将上面所建模型应用到不同城市(选择南方和北方共 10 个城市)，取上述城市最近 2000~2020 年的降水量数据，预测 2021 年的降水量。计算显示，在预测降水量状态方面，除了杭州、南京的错了以外，别的城市的降水量状态都预测准确。在用模糊集理论预测降雨量方面(见表 9)，相对误差超过 20%的占 20%，相对误差在 20%内的占 80%，相对误差在 10%内的占 70%，总体预测效果较好。南方的城市相对来说误差比较高，有的甚至预测失败，这说明所建模型还有需要改进的地方。

Table 9. Predicted values for different cities
表 9. 不同城市的预测值

城市	真实值	预测值	相对误差	城市	真实值	预测值	相对误差
南通	1284.4	1159.4426	9.7%	**嘉兴**	1606.9	1364.3443	15.1%
杭州	1929.9	1423.6608	26.2%	**常州**	1142.6	1069.4983	6.4%
无锡	1142	1060.1985	1.6%	**北京**	698.4	745.0412	6.7%
石家庄	861.5	873.6370	1.4%	**哈尔滨**	640.8	670.4828	4.8%
南京	1267.1	1006.4478	20.6%	**长沙**	1472.983	1546.7810	5.0%

7. 结语

加权马尔可夫链对降水量的预测北方地区的准确率要高于南方[9] [10]。上海地区属于南方，降水量大，本文根据上海的降雨量特点做了有针对性的分组，并在大量计算的基础上对模糊集理论的公式做了一些修改以使得预测结果更加准确。结果显示平均误差为4.5%，预测精度较高。将所建模型应用到不同城市，结果显示大部分城市预测结果较好，但有部分城市预测失败。这说明模型有需要提高的地方，这是后面继续研究要解决的问题。

基金项目

本文由上海工程技术大学大学生创新项目：螺纹钢期货价格影响因素分析(项目编号：CS2221003)资助。

参考文献

[1] 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+98.

[2] 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.

[3] 孙才志, 林学钰. 降水预测的模糊权马尔可夫模型及应用[J]. 系统工程学报, 2003, 18(4): 294-299.

[4] 李娟, 张维江, 马轶. 滑动平均——马尔可夫模型在降水预测中的应用[J]. 水土保持研究, 2005, 12(6): 196-198.

[5] 王涛, 钱会, 李培月. 加权马尔科夫链在银川地区降水量预测中的应用[J]. 南水北调与水利科技, 2010, 8(1): 78-81.

[6] 张杰, 陶望雄, 王青. 加权马尔科夫链在济南市降水量预测中的应用[J]. 人民黄河, 2016, 38(9): 13-16.

[7] 李亚斌, 徐盼盼, 钱会, 等. 加权马尔可夫链在铜川地区降水量预测中的应用[J]. 灌溉排水学报, 2017, 36(5): 96-102.

[8] 梁显丽, 宝秋利, 代海燕. 加权马尔可夫链在鄂尔多斯市年降水量预测中的应用[J]. 数学的实践与认识, 2021, 51(4): 161-171.

[9] 宋帆, 杨晓华, 武翡翡, 等. 基于聚类分析的模糊马尔科夫链在降雨量预测中的应用[J]. 节水灌溉, 2018(10): 33-36+41.

[10] 董亚, 杜景林, 胡玉杰. 基于K-means的模糊马尔科夫模型的降水预测[J]. 现代电子技术, 2021, 44(19): 85-89.