基于非参数核密度估计模型的乌鲁木齐河月径流随机模拟

陈大春，何 英，曹 伟

（新疆农业大学水利与土木工程学院，新疆 乌鲁木齐 830052）

摘 要：利用非参数核密度估计方法建立了乌鲁木齐河月径流随机模拟的NP模型。其中，通过以最小二乘交叉验证(LSCV)指标为目标的粒子群优化获取NP模型带宽参数；采用可变核带宽方法进行边界修正。使用1958~2010年间53a月径流数据，经过250组分组模拟进行实用性检验。最后，与使用SAMS2007所建立的季节自回归PAR模型进行了对比。结果表明：所建乌鲁木齐河月径流NP模型能较好保持原序列统计特性；与PAR模型相比，它具有参数少、计算简单的特点。

关键词：随机模拟；核密度估计；非参数；乌鲁木齐河

中图分类号：TV121⁺.4 P333.9 **文献标识码**：A **文章编号**：1000-0852(2014)02-0066-05

在水资源规划、设计和管理中，水文时间序列模型是揭示、重构实测水文序列统计特征的重要方法。随着随机水文学的发展，水文时间序列模型发展为参数模型和非参数模型。参数模型如ARIMA、解集模型、散粒噪声模型、门限自回归模型等均是根据经验对实测数据的概率分布和相依性等做了一定简化后建立起来的。它们虽然相对简单、应用方便，能基本表征水文时间序列的统计特征和一般变化规律，但是这些简化却难以全面描述水文现象的客观规律。而非参数模型是建立在数据驱动基础上的，它的优势在于可以不对研究对象的概率分布和相依形式作相关假定，实用性更强。在乌鲁木齐市水资源短缺风险分析中，需要使用水文随机模拟方法产生人工月径流序列。

水文非参数随机模拟方法主要有两类：一类是由Sharma等[1]提出的基于径流序列的马尔科夫性假设的NP模型（Nonparametric order p Markov Streamflow Model）；另一类则是Lall和Sharma[2]提出的k-NN模型（k Nearest Neighbor Resampling Model）。1997年Sharma等[1]提出了基于径流序列的马尔科夫性假设的NP模型。NP模型中，使用核函数方法进行概率密度估计并推导出条件概率密度估计方法。依据条件概率由各单变量密度估计核函数的加权和构成这一特点，以分布权重为抽样概率选择状态间联合分布并计算出均值、方差，并由此计算出模拟值。由于高斯核会使模拟值出现边界问题，NP模型需进行重复抽样并造成部分概率密度函数失真。此后，水文非参数随机模拟模型得到广泛推广及改进。Tarboton[3]将NP模型进一步推广到解集模型并形成了非参数解集模型NPD（Nonparametric Disaggregation Model）。2002年，Sharma和O′neill[4]在非参数估计框架下为解决大时滞（季节到年际）相依结构模拟存在的问题，使用可变核（variable kernel）和径流聚集变量（aggregate streamflow variable）改进了NP模型。在Beaver河月径流和Burrendong坝入流应用中与NP模型相比，提高了对长时相依结构的模拟精度。王文圣等[5]将NP模型扩展为多变量非参数模型并成功应用到金沙江的屏山站和宜宾站的日径流模拟中。

本文使用乌鲁木齐河英雄桥站1958~2010年的月径流资料，利用粒子群优化算法（Particle Swarm Optimization，PSO）计算各月核密度带宽，建立了乌鲁木齐河月径流NP随机模拟模型。与季节自回归模型（PAR）模型进行对比验证结果表明，所建乌鲁木齐河月径流NP模型可靠可行。

收稿日期：2013-06-25
基金项目：新疆高校科研计划重点项目（XJEDU2011I22）
作者简介：陈大春（1973-），男，重庆人，硕士，副教授，研究方向为水资源规划与管理。E-mail：vision_studio@163.com

1　研究区域及资料

乌鲁木齐河流域位于天山北坡中段，东经86°45′~87°56′，北纬43°00′~44°07′之间。西接头屯河流域，东为板房沟流域，流域总面积4 684km²，其中山区（西白杨沟口以上）流域面积1 070km²，流域平均海拔3 006m；英雄桥以上流域面积924km²，平均海拔3 083m。乌鲁木齐河属降水和冰雪融水补给型河流，英雄桥水文站连续最大四个月径流量出现在6~9月，占年径流量的78.7%。最大月径流与最小月径流之比为26.8，见图1。

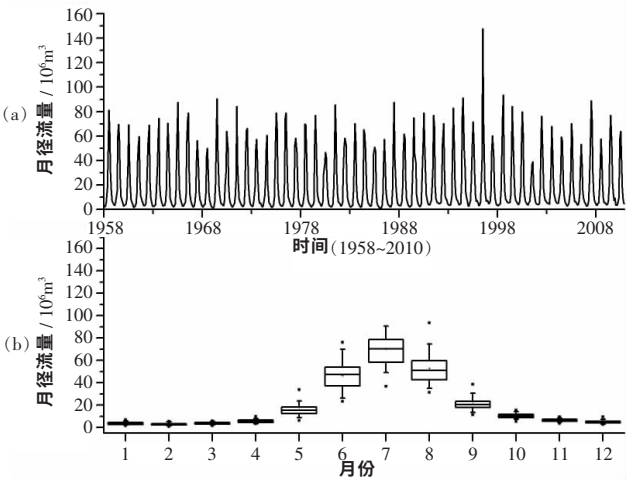

图1　乌鲁木齐河月径流量(1958~2010)
(a)年季变化折线图(b)季节变化箱图

Fig. 1 Monthly streamflow of the Urumqi River(1958–2010)
(a)Line chart of annual variation; (b)Box plot of seasonal variation

2　NP模型

2.1　模型原理

核密度估计方法在水文时间序列模型中的应用首次由Sharma[1]提出，并命名为非参数p阶马尔科夫径流模型NP_p (Nonparametric Order p Markov Streamflow Model)。

核密度估计由多个样本数据经验频率分布的加权移动平均组成。样本的核密度估计定义为[6]：

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right) \tag{1}$$

式中：$K(\ )$为核函数，为指定概率密度函数，即$\int_{-\infty}^{+\infty}K(x)dx=1$；正数$h$被称为带宽(bandwidth)，决定核密度估计的光滑程度。

多数研究表明：与非参数核回归相同，相对于带宽h的选择核K的选择并不重要。通常情况下均选择高斯核函数，d维向量x的高斯核密度函数描述如下：

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(2\pi)^{d/2}\det(H)^{1/2}}\cdot\exp\left(-\frac{(x-x_i)^TH^{-1}(x-x_i)}{2}\right) \tag{2}$$

式中：n为实测向量x的个数；H为$d\times d$维带宽阵（对称正定矩阵）。H由带宽h和数据样本协方差阵计算而来：

$$H=h^2S \tag{3}$$

经过推导后，单变量条件概率密度函数为[7]：

$$\hat{f}(x_t|V_t)=\sum_{i=1}^{n}W_i\frac{1}{(2\pi)^{1/2}\det(c)^{1/2}}\cdot\exp\left[-\frac{(x_t-b_i)^2}{2c}\right] \tag{4}$$

而

$$W_i=\exp\left\{-\frac{(V_t-V_i)^TS_V^{-1}(V_t-V_i)}{2h^2}\right\}\Big/\sum_{j=1}^{n}\exp\left\{-\frac{(V_t-V_j)^TS_V^{-1}(V_t-V_j)}{2h^2}\right\} \tag{5}$$

$$\sum_{i=1}^{n}W_i=1\quad b_i=x_i+(V_t-V_i)^TS_V^{-1}S_{xV}^T\quad c=h^2(S_x-S_{xV}S_V^{-1}S_{xV}^T)$$

式中：$V_t=(x_{t-1},x_{t-2},\cdots,x_{t-p})^T$为测量值$x_t$的前$p$个值；$S_x$，$S_{xV}$，$S_V$为样本协方差矩阵的各子矩阵。

$$S=\begin{pmatrix}S_x & S_{xV}\\ S_{xV}^T & S_V\end{pmatrix} \tag{6}$$

从式(4)中可以看出条件密度函数是$n-p$个高斯函数的加权平均，其中b_i和c分别为均值和方差。

使用下式进行随机模拟：

$$x_t=b_i+\sqrt{c}\ e_t \tag{7}$$

式中：e_t为均值0、方差1的高斯随机变量。

与回归随机模拟方法相似，因为使用的高斯核函数具有对称无界特点，模拟结果存在边界问题。常用的核密度边界修正方法有：边界核、反射、变换和局部多项式拟合等。本文采用可变核方法以减小边界影响。

当计算式(7)时，使用可变核函数带宽以计算方差c：

$$h'=\begin{cases}h, & F_{N(b_i,h^2S')}(x_t\le 0)\le\alpha\\ F_{N(b_i,h'^2S')}(x_t\le 0)=\alpha, & F_{N(b_i,h^2S')}(x_t\le 0)>\alpha\end{cases} \tag{8}$$

式中：h'为边界修正带宽；$F_{N(b_i,h^2S')}$为高斯分布的累积分布函数；α为阈值。

从式(8)看出，当估计的径流小于等于0的概率大于阈值时，计算阈值α所对应的高斯分布方差并以此计算出核函数带宽h'。因此，当用式(7)模拟月径流x_t

前,先用式(8)计算出可变核函数带宽 h',并重复计算直至得到大于零的 x_t 值。本文中 α 取 0.05。

2.2 模拟步骤

模拟步骤见图 2。

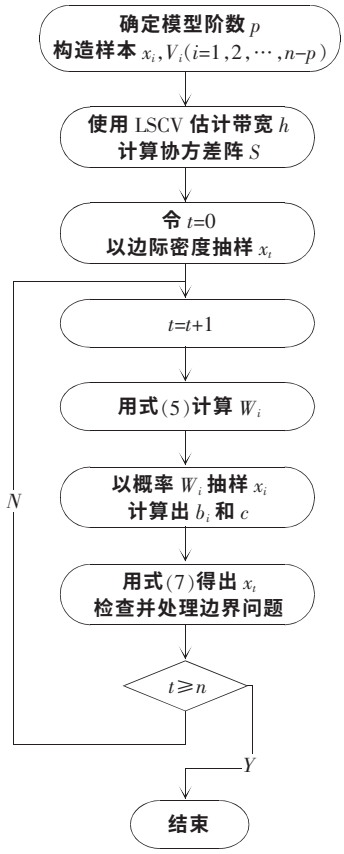

图 2 NP 非参数随机模拟步骤

Fig. 2 The simulation step of the nonparametric stochastic

2.3 乌鲁木齐河月径流 NP 模型参数

(1)模型阶数 p

模型阶数一般使用自相关分析初步确定后,以 AIC(Akaike Information Criterion)准则进行识别。AIC 准则为:

$$AIC(p)=2p+n\ln\sigma_\varepsilon^2 \tag{9}$$

式中:p 为模型阶数;n 为序列长度;σ_ε^2 为残差的方差。

对原月径流序列进行 X-12 季节调整,调整后的自相关函数 ACF 和偏自相关函数 PACF 如图 3。

从 PACF 图看出自回归阶数为 2 或 5,但根据表 1 计算的 AIC 结果选择模型阶数为 2 阶较合适。

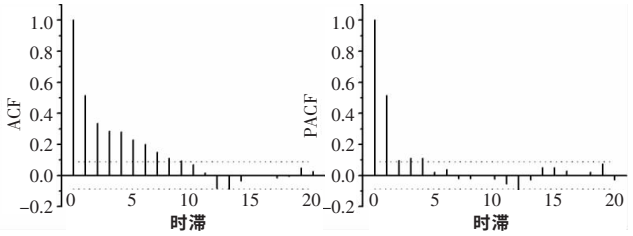

图 3 季节调整后的 ACF 和 PACF

Fig. 3 The seasonally adjusted ACF and PACF

表 1 乌鲁木齐河月径流序列各阶 AIC 值

Table 1 AIC values of the monthly runoff series of the Urumqi River

p	1	2	3	4	5
AIC	5.5812	5.5698	5.5701	5.5746	5.5702

(2)NP 模型带宽 h

核密度估计中的一个难点是合适带宽 h 的确定,小的带宽给出很粗糙的估计,而大的带宽给出较光滑的估计。带宽估计方法主要有如:均方误(MSE)、误差平方积分(ISE)和最小二乘交叉验证(LSCV)等指标方法及嵌入带宽(plug-in bandwidth)和适应性核(adaptive kernel)等方法[8]。本文使用 LSCV 作为带宽分析指标,其计算方法如式(9):

$$LSCV(H)=\frac{1+\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}[\exp(-L_{ij}/4)-2^{p/2+1}\exp(-L_{ij}/2)]}{(2\pi^{1/2})^p n\det(H)^{1/2}} \tag{10}$$

式中:H 由带宽 h 和数据样本协方差阵 $H=h^2S$ 计算而来;x_i, x_j 是不同数据;L 用下式计算:

$$L_{ij}=(x_i-x_j)^T H^{-1}(x_i-x_j) \tag{11}$$

推荐带宽 h 的取值范围为 0.25~1.1 倍 h_{ref},本文采用范围为 0.25~1.3 倍 h_{ref}。其中 h_{ref} 为:

$$h_{ref}=\left(\frac{4}{p+2}\right)^{1/(p+4)} n^{-1/(p+4)} \tag{12}$$

式中:p 为模型阶数;n 为数据数。

用参数优化方法最小化 LSCV 得分可以得到合适的核函数带宽。本文使用粒子群优化算法优化得到乌鲁木齐河月径流序列各月带宽如表 2。

表 2 乌鲁木齐河月径流序列 NP_2 核密度估计各月优化带宽 h

Table 2 The optimized bandwidth h for the NP_2 kernel density of monthly runoff series of the Urumqi River

1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
0.438	0.468	0.525	0.581	0.153	0.636	0.514	0.627	0.349	0.254	0.621	0.430

3　模型检验及分析

本文使用乌鲁木齐河英雄桥站1958~2010年间53年月径流资料作为建模数据。将月径流资料$\{z_{i,j}, i=1,2,\cdots,53; j=1,2,\cdots,12\}$按$j=12$个截口分别形成$x_i, V_i$:

$j=1$时,$V_i=(z_{i-1,12}, z_{i-1,11}, \cdots\cdots, z_{i-1,j-p+1})$,$x_i=z_{i,1}$

$j=2$时,$V_i=(z_{i,1}, z_{i-1,12}, \cdots\cdots, z_{i-1,j-p+2})$,$x_i=z_{i,2}$

…　　…

$j=p$时,$V_i=(z_{i,p-1}, z_{i,p-2}, \cdots\cdots, z_{i-1,12})$,$x_i=z_{i,p}$

为检验模型实用性,产生250组样本,每组样本容量为52a。使用美国科罗拉多州立大学研制的SAMS 2007软件建立乌鲁木齐河月径流PAR模型并与NP模型进行对比。在SAMS中对原序列进行边际概率偏态分布处理,经检验2、3和6月采用Gamma分布,其它月份采用对数正态分布。

为对比显示模拟结果,本文分别绘制了有代表性的3、7和10月人工径流的边际概率密度图及均值、标准差、偏度系数、1~2阶相关系数等统计参数图,见图4。由于数据量大,本文以箱图绘制模拟结果,其中一个盒子代表一个统计参数的250个估计值的分布状态。上下盒须分别代表5%和95%分位数,中间线为均值。而图中实折线代表原历史资料的统计结果。

(1)边际概率密度

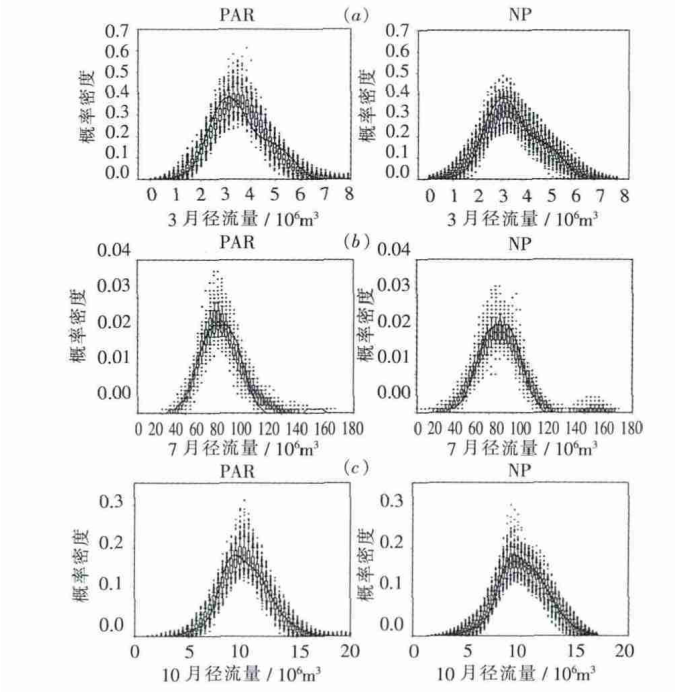

图4　(a)3月、(b)7月和(c)10月模型模拟的边际概率密度箱图对比

Fig. 4 Box plots of marginal density estimates for (a)March, (b)July and (c)October

为检验人工径流序列是否服从原序列分布,常见方法是检验人工序列的统计参数是否偏离原序列的统计值或进行拟合优度检验,而最直观的方法是对比检验人工序列的密度函数形态。图4中实折线为历史数据各月径流概率密度曲线。从图4看出,SAMS的PAR(2)模型无法模拟出多峰概率密度函数并有一定偏差,而NP(2)模型基本保持了原密度函数的形状。由于使用抽样选取某一片高斯密度函数来模拟真实条件密度函数而导致模拟序列密度函数均有一定程度低估现象,这是此方法需要改进之处。

(2)统计参数

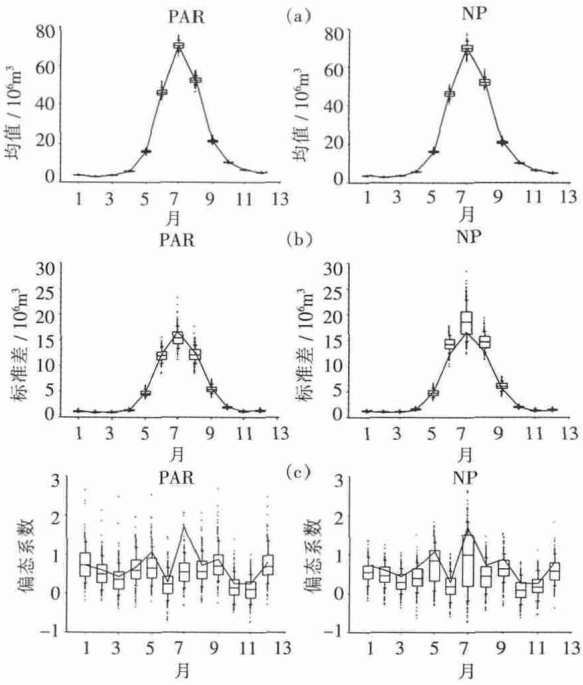

图5　PAR和NP模型的(a)均值、(b)标准差和(c)偏度箱图

Fig. 5 Box plots of (a)mean, (b)standard deviation and (c)skewness of PAR and NP models

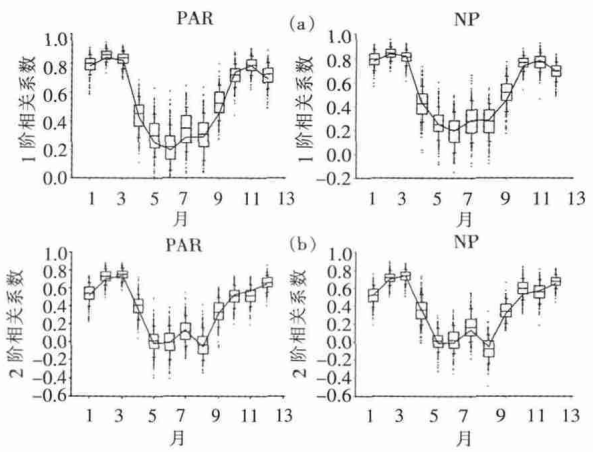

图6　PAR和NP模型的(a)1阶、(b)2阶相关系数

Fig. 6 (a)Lag 1 and (b)lag 2 correlations of PAR and NP models

从统计参数看(见图5、图6),两模型的均值均能很好保持,而NP模型的标准差偏大($\sigma^2=(1+h^2)S_x$,见文献[1])。偏度估计中,PAR模型在多数月份中出现低估且多离群点,而NP模型在5、6、7月份的模拟情况要较PAR模型要好。1阶相关系数估计中,NP模型能完全保持原序列相关特性,2阶相关系数估计两模型则效果相当。

4 结语

本文使用核密度估计方法建立了乌鲁木齐河月径流非参数随机模拟模型并与传统的季节自回归模型PAR进行了比较。结果表明,所建的乌鲁木齐河NP随机模拟模型能较好保持原序列的概率分布和统计参数。与PAR模型相比,NP模型需估计的参数较少,不需对原序列进行转换及平稳性检验。不足之处在于,核密度函数的带宽估计较复杂,尤其本文分别对各月取不同带宽;边界问题尚待进一步研究。

参考文献:

[1] Sharma A, Tarboton D G, Lall U. Streamflow simulation: a non-parametric approach [J]. Water Resources Research, 1997, 33(2): 291-308.

[2] Lall U, Sharma A. A nearest neighbor bootstrap for resampling hydrologic time series [J]. Water Resources Research, 1996, 32(3): 679-693.

[3] Tarboton D G, Sharma A, Lall U. Disaggregation procedures for stochastic hydrology based on nonparametric density estimation [J]. Water Resources Research, 1998, 34(1): 107-119.

[4] Sharma A, O´neill R. A nonparametric approach for representing interannual dependence in monthly streamflow sequences [J]. Water Resources Research, 2002, 38(7): 5.1-5.10.

[5] Wang W, Ding J. A multivariate non-parametric model for synthetic generation of daily streamflow [J]. Hydrological Processes, 2007, 21(13): 1764-1771.

[6] Silverman B W. Density Estimation for Statistics and Data Analysis [M]. Chapman & Hall/CRC, 1986.

[7] 王文圣, 丁晶. 单变量核密度估计模型及其在径流随机模拟中的应用 [J]. 水科学进展, 2001, 12 (3):367-372. (WANG Wensheng, DING Jing. Kernel density estimation model and its application to stochastic generation in hydrology and water resources [J]. Advances in Water Science, 2001,12(3):367-372. (in Chinese))

[8] Sharma A, Lall U, Tarboton D. Kernel bandwidth selection for a first order nonparametric streamflow simulation model [J]. Stochastic Hydrology and Hydraulics, 1998,12(1):33-52.

Urumqi River Monthly Runoff Stochastic Simulation Based on Non-parameter Kernel Density Estimation Model

CHEN Dachun, He Ying, Cao Wei

(1. *College of Water Conservancy and Civil Engineering, Xinjiang Agricultural University, Urumqi* 830052, *China*)

Abstract: By using non-parametric kernel density estimation method, a NP stochastic simulation model of Urumqi River monthly runoff was established. Moreover, a particle swarm optimization model using a LSCV (Least Squares Cross-Validation) as the objective function was employed to obtain bandwidth parameters; a variable kernel bandwidth method was adopted to adjust kernel density boundary. Afterwards, 53-year (1958-2010) records of monthly runoff were applied for 250 simulations, each with a length of 53 years, were made to carry on the practicability test of the model. Finally, the results from a PAR (seasonal autoregressive) model built by software SAMS2007 was presented for comparison of NP model. The results show that the Urumqi River monthly runoff NP model can better maintain the statistical properties of the original sequence. Comparing with the PAR model, NP model has the characteristics of fewer parameters and simple calculation.

Key words: stochastic simulation; kernel density estimation; non-parameter; Urumqi River