

非参数统计在洪水频率分析中的应用与展望

董洁¹, 谢悦波¹, 翟金波²

(1. 河海大学水资源环境学院, 江苏 南京 210098; 2. 泰安市规划局, 山东 泰安 271000)

摘要: 介绍非参数统计方法在洪水频率分析中的应用现状及其进展, 指出把地区信息和历史洪水信息纳入密度函数估计中、把水文的模糊性和灰色性纳入非参数统计模型中以及解决洪水频率曲线的外延等等, 是非参数统计在洪水频率分析领域的发展方向。

关键词: 非参数统计; 核函数估计; 洪水频率分析; 窗宽

中图分类号: P333.9 **文献标识码:** A **文章编号:** 1000-1980(2004)01-0023-04

在各种水利规划设计中, 都需要进行水文频率分析工作, 以保证达到工程要求的设计值。解决这一问题通常用参数统计方法。在参数统计方法中, 线型的选择正确与否是至关重要的, 它直接影响着设计值的精度。然而, 在水文频率计算中, 所使用的线型只是一种推断, 与实际分布很难吻合, 因而参数统计方法的精度也受到很大的限制。近代非参数统计的发展, 为水文频率计算提供了另一条研究途径。它避开了水文频率计算中困惑多年的线型问题, 不需要线型假设的先决条件, 而是直接由实测系列与历史洪水求出较合理的设计值, 从而克服了参数统计方法中模型选择的主观性, 较真实地反映了水文水资源系统的客观规律。应用研究表明, 非参数统计方法在水文频率计算中具有很大的潜力。本文对非参数统计方法在水文频率计算中的应用和研究现状进行介绍, 并展望了今后的研究和发展方向。

1 应用现状

1.1 几种重要的密度估计法

1.1.1 直方图法

直方图法基于概率密度的一个基本性质: 如有 X 的简单样本 $X_1, \dots, X_n$, 则概率 $P(a \leq X \leq b)$ 可用 $\# \{i: 1 \leq i \leq n, a \leq X_i \leq b\} / n$ 去估计。因此, 当 $b - a$ 充分小时, $\# \{i: 1 \leq i \leq n, a \leq X_i \leq b\} / n(b - a)$ 可近似代表 $f(x)$ 在区间 $[a, b]$ 上的值。这样就得到 $f(x)$ 的一个估计值。基于上述原理, 可选择适当的正数 h , 把全直线分成一些长为 h 的区间, 任取这些区间之一, 记为 I 。对 $x \in I$ 以 $\# \{i: 1 \leq i \leq n, X_i \in I\} / (nh)$ 作为 $f(x)$ 的估计。重要的是 h (一般称为窗宽) 的选择。 h 太大, 平均化的作用突出而淹没了密度的细节部分; h 太小, 则随机性的影响太大, 从而产生极不规则的图形。直方图法有一个致命的缺点: 对每个区间边缘部分密度的估计较差。Rosenblatt 对此进行了改进。

1.1.2 Rosenblatt 法

指定一个正数 h 如前, 对每一个 x , 记 I 为 x 的中心, 长为 h 的区间, 即 $[(x - h)/2, (x + h)/2]$ 。算出之值作为 f 在点 x 处的值 $f(x)$ 的估计。Rosenblatt 法事先不把分割区间定下, 而是让区间随着估计点 x 跑, 使 x 始终处在区间的中心位置, 从而获得较好的效果。

1.1.3 核估计

Rosenblatt 法虽然比直方图法好, 但其仍为一阶梯函数, 不是连续函数。1962 年 Parzen 提出了核估计法。

设 $k(\cdot)$ 为给定的概率密度函数, $h > 0$ 是一个与 n 有关的常数, 定义 $\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n k\left(\frac{x - x_i}{h}\right)$ 为总体未知密度 f 的一个核估计。其中 $k(\cdot)$ 为核函数, h 为窗宽。目前密度估计法中核估计在理论上发展最完善, 本文中将其重点讨论。核估计既与样本有关, 又与核函数及窗宽的选取有关。在给定样本以后, 一个核估计

的好坏,取决于核及窗宽的选取是否得当.核函数和窗宽的选择直接影响设计值的精度,因而引起了许多学者的关注和讨论,下面介绍这方面已取得的一些成就.

1.2 核函数和窗宽的选择

Rosenblatt 和 Parzen 分别于 1956 年和 1962 年提出了非参数概率密度估计法——核估计,随后许多人也投入到这方面的工作中,他们是 Scott^[1],Bowman^[2],Terrell, Silverman, Sheater, Marron, Wand 等.

1.2.1 核函数的选择

从理论上来说,不一定要要求核函数为密度函数,但从实用上要求核函数 $k(\cdot)$ 为密度函数是合适的.因为待估计的函数 $f(x)$ 也是密度函数,所以估计量最好是密度函数.原则上可对核函数 $k(\cdot)$ 施加一定的限制,使估计量与待估函数偏差在一定意义上尽可能地小.例如可要求核函数 $k(\cdot)$ 具有对称性,其一阶矩为零,且有有界性、连续性等等.任何概率密度函数都可以作为核函数,但是用于水文频率分析的核函数更需重视其尾部特征,并使积分均方差最小^[3].因此,优选核函数一般基于以下几方面考虑.

1.2.1.1 基于积分均方差的考虑

Scott 等^[1]将积分均方差定义为 $MISE = \int E(\hat{f}(x) - f(x))^2 dx$,并给出了理想核函数,如矩形核(rectangular)和 Cauchy 核等.从理论上讲,用于洪水频率分析的核函数一般要求具有非负性^[1]、对称性^[1]、最小 MISE 等等.但是在实际应用中,许多学者建议放宽对称性的要求,使其更符合水文频率分析的特性.

1.2.1.2 基于核函数尾部特征的考虑

所选用核函数的尾部特征及其曲线的光滑性,将直接影响该核函数的密度估计,而且在洪水频率分析中人们感兴趣的是推断超过实测系列的洪峰流量.因此,许多具有尾部特征的核函数被提出来:Adamowski 等^[3]提出 Gaussian, Cauchy, EV1 核.

1.2.1.3 基于核函数形状上的考虑

一般较优核函数要求具有对称性,因为这能够使密度估计无偏, Gaussian 核和 Cauchy 核是对称的, EV1 是非对称的.在洪水频率分析中,洪水系列一般有较小的下边界,且密度函数是非对称的.因为非对称密度函数经过对数变换后很容易变为对称的,所以,在实际应用中可放宽对称性的要求.

1.2.2 窗宽的选择

一般认为,窗宽的选择比核函数的选择更重要.因为选择窗宽的目的是为了进行核的密度估计.窗宽一般随着样本容量 n 的增大而下降.窗宽取得太小时,随机性的影响增加,而 $\hat{f}(x)$ 呈现不规则的形状,这可能会掩盖 $f(x)$ 的重要特征;反之,窗宽太大则 $\hat{f}(x)$ 将受到过度的平均化,使 $f(x)$ 比较细致的性质不能显露出来.

1.2.2.1 固定核的方法(FKE)

对于来自某一总体的样本,在进行核估计时,选择一个理想的窗宽的标准是使 MISE 最小.由于 $f(x)$ 未知,所以窗宽一般是从经验公式得出, Adamowski^[3], Scott and factor^[1], 以及 Labatiuk 和 Adamowski 在这方面进行了研究.固定核方法中的窗宽一般在用样本进行密度估计时保持不变,因此在样本点集中的部分窗宽显得太大,使曲线过于光滑,而在样本较稀疏的部分窗宽又显得太小而导致曲线不够光滑.所以固定核估计法对于处理有长尾和多峰的情况不够灵活,因此给出变核估计法.

1.2.2.2 变核估计法(VKE)

Breiman 等^[4]给出变核定义为 $\hat{f}(x) = \frac{1}{N} \sum_{j=1}^N \frac{1}{a_k d_{j,k}} k\left(\frac{x - x_j}{a_k d_{j,k}}\right)$,其中 $d_{j,k}$ 是样本中 x_j 的第 k 个最邻近点的距离, a_k 是连续的光滑因子.大量模拟研究证明,变核估计法优于固定核法. Breiman 等^[4], Habbema 等^[5], Raatgever 等^[6], Bean 等^[7], Bowman^[2], Adamowski^[8]都对变核估计法进行了研究.除此之外,窗宽的选择方法还有自适应窗宽优选法、数字跟踪法等.

以上是对核估计方法方面的讨论,下面介绍对核估计法在水文频率分析方面的进展情况.

1.3 核估计法在洪水频率分析中的应用

1985 年 Adamowski 首次将核密度估计方法应用于洪水频率分析^[3],其以实测资料为依据,把非参数核密度估计法与不同假设条件下参数统计模型(LN3, LPⅢ, EV)的计算结果进行分析,并得出结论:核估计法比上

述参数统计模型具有较小的偏差和均方差,并且该法能解决洪水的多峰问题,这是参数方法所不能及的。

1.4 参数和非参数相结合的洪水频率分析

由于非参数核估计曲线外延有限,即非参数估计模型对超出样本范围许多的稀遇频率的设计值无法得到,且不能充分利用洪水信息,因此 Schuster 等^[9]提出了一种参数与非参数混合的密度估计,并证明了它的有效性。这种方法虽然把地区信息及经验分析加入模型当中,同时也解决了密度曲线的尾部问题,但是仍然不能回避线型的选择问题。

1.5 考虑历史洪水和古洪水的非参数密度估计

郭生练^[10]在考虑我国实测资料比较短而古洪水研究又很有特色的情况下,提出考虑历史和古洪水的非参数方法,并通过实例与参数模型进行了对比研究,得出非参数方法的描述和预测能力皆优于参数方法的结论。

1.6 非参数核密度估计结合再抽样方法的洪水频率分析

Tung 等^[11]首次将非参数法——自助法(或称再抽样法)应用于洪水频率分析。此法主要是计算参数的样本方差。夏乐天^[12]用非参数方法结合再抽样法对核估计进行了纠偏处理,并用统计试验方法在 P-Ⅲ型总体与三参数对数正态总体的情况下对非参数密度估计法及各总体的最优参数估计的稳健性进行了比较,得出的结论是非参数法的稳健性最好。

2 非参数方法在水文频率分析领域的应用展望

非参数模型能拟合各种实测数据资料,包括多峰分布的数值,而参数模型则做不到。与参数模型方法相比,非参数统计方法在水文频率分析中的应用研究还处于初步研究阶段,有许多问题还需要更深入地进行研究。

2.1 在理论方面

目前,非参数理论虽有很大的发展,但是还很不成熟。比如针对核函数和窗宽的选择,虽然在理论上进行了大量的讨论,也确定了一些方法和原则,但没有统一的标准确定哪一种核函数是较优的,在实际应用中可操作性差,针对不同的问题如何选择还有待研究。另外,非参数方法是一种通过样本数据直接估计的方法,因而具有较强的适应性,但是它不能把地区信息、洪水的不确定信息加入到非参数模型中,洪水频率曲线的外延也存在许多不足。

2.2 在实际应用方面

在洪水频率分析中,人们期望所选择的模型具有以下几个方面的特点:(a)细致地描述水文特征,比如多峰情况;(b)对不同区域能够方便地采用统一的模型计算;(c)模型中能包含更多的地区域信息;(d)模型要具有一定的尾部特征并由此推求界外值,且能得到稳定的结果。非参数方法能够很好地解决前两个问题,关于后两个问题目前虽然已经取得了一些成果,但还不够深入,需要做进一步的研究。例如,为了解决曲线的外延问题,要求选择一个非对称的核函数,但是从理论上讲,对称的核函数可以使偏差和均方差最小。因此,如何协调理论与应用之间的矛盾也是亟待解决的问题。

总之,如何把地区信息和历史洪水信息纳入密度函数估计中,如何把水文的模糊性和灰色性纳入非参数统计模型中,如何解决曲线的外延问题,等等,都将是今后研究的努力方向。可以预计,随着非参数统计理论研究的不断完善和发展,它定会在水文领域中得到更加广泛的应用。

参考文献:

- [1] SCOTT D W, FACTOR L E. Mont Carlo study of the data based nonparametric probability density estimators[J]. Am Stat Assoc, 1998, (80):9—15.
- [2] BOWMAN A W. A comparative study of some kernel-based nonparametric density estimators[J]. Stat Comput simul, 1985, 21:313—327.
- [3] ADAMOWSKI K. Nonparametric kernel estimation of frequencies[J]. Water Resources Research, 1985, 21(11):1585—1590.
- [4] BREIMAN L W M, PURCELL E. Variable kernel estimates of multivariate densities[J]. Technometrics, 1997, 19(2):114—135.

- [5] HABBEMA J D F, RRMME J. Variable kernel density estimation in discriminant analysis in compstan[J]. Physica Verlay Vienna, 1978, (2): 178—185.
- [6] RAATGEVER J W, DUIN R P W. On the variable kernel model for multivariate nonparametric density estimation[J]. Physica Verlay Vienna, 1978, (3): 524—533.
- [7] BEAN S, TSOKOS C P. Bandwidth selection procedures for kernel density estimation[J]. Theory Methods, 1982, 11: 1045—1069.
- [8] ADAMOWSKI K. A Mont-Carlo comparison of parametric and nonparametric estimation of flood frequencies[J]. Journal of Hydrology, 1989, 108: 295—308.
- [9] SCHUSTER E, YAKOWITZ S. Parametric/Nonparametric mixture density estimation with application to flood-frequency analysis[J]. Water Resources Bulletin, 1985, 21(5): 797—814.
- [10] GUO Sheng-lian. Nonparametric variable kernel estimation with historical floods and paleoflood information[J]. Water Resources Research, 1991, 27(1): 91—98.
- [11] TUNG Y K, MAY L W. Reducing hydrologic parameter uncertainty[J]. ASCE(WR1), 1981, 1: 123—142.
- [12] 夏乐天. 水文频率计算中的一种非参数估计方法——密度函数估计法[A]. 见: 国际水文计划中国国家委员会主编. 2000年中国水文展望[C]. 南京: 河海大学出版社, 1991. 86—92.
- [13] 陈希儒, 柴根象. 非参数统计教程[M]. 上海: 华东师范大学出版社, 1993. 247—287.
- [14] 王文圣, 丁晶. 非参数统计方法在水文水资源中的应用与展望[J]. 水科学进展, 1999, 10(4): 458—463.
- [15] ADAMOWSKI, FELUCH W. Nonparametric flood frequency analysis with historical information[J]. Hydraul Eng, 1990, 116(8): 1035—1047.
- [16] KYANG J. Comparative study of flood quantities estimation by nonparametric modes[J]. Journal of Hydrology, 2002, 260: 176—193.
- [17] YOUNG M, VPMANU L. A comparison of tail probability estimators for flood frequency analysis[J]. Journal of Hydrology, 1993, 151: 343—363.
- [18] UPMANU L, UOUNG M. Kernel flood frequency estimators: Bandwidth selection and kernel choice[J]. Water Resource Research, 1993, 29: 152—156.
- [19] KITE G W. Frequency and risk analysis in hydrology[M]. Fort Collins Colo: Water Resources Publications, 1977. 145—156.

Application of nonparametric statistic approach to flood frequency analysis and prospect of its development trend

DONG Jie¹, XIE Yue-bo¹, ZHAI Jin-bo²

(1. College of Water Resources and Environment, Hohai Univ., Nanjing 210098, China;

2. Taian Bureau of Planning, Taian 271000, China)

Abstract: By analysis of the current situation of and some advances in the application of the nonparametric statistic approach to flood frequency analysis, it is pointed out that the introduction of the regional and historical flood data into the density function estimation and grey and fuzzy attributes of hydrological statistics into the nonparametric statistic model, and further extension of the flood frequency curve are the development trend of the nonparametric statistic approach in the field of flood frequency analysis.

Key words: nonparametric statistics; kernel function estimation; flood frequency analysis; bandwidth