直径结构模拟中的核方法与直方图及列点法的比较分析*

王雪峰　唐守正

摘要　应用非参数核密度估计方法,可以由样本直径结构很好地描述总体直径结构,而无需假定总体的直径分布。本文采用计算机模拟技术,比较了核方法与直方图方法及列点法在描述总体时的优劣,结论是核方法优于直方图方法,也优于列点法。

关键词　模拟、非参数核密度估计、直径分布

对直径结构的研究已有近百年的历史,从研究者所使用的方法看,主要有列点法(List of diameters),林分表法(Stand table),分布函数法(Distribution Function)和百分位法(Percentile)。在早期的直径结构研究中,主要采用前两种方法,即着重从生物学角度研究林木直径大小序列,采用简单的统计数据,以列点法或直方图(Histogram)研究直径结构规律。以后计算机的发展,给人们提供了求解复杂函数的可能性,故人们把重点转移到对分布函数的研究。较典型的分布函数有对数正态分布[1]、Γ 分布[2]、β 分布[3]、Sb 分布[4]、Weibull 分布[5]。其共同点是都需要知道林分直径的具体分布形式。从现有研究中各种分布函数的拟合结果看,无论假定哪一种分布,卡方检验的接受率都很低[1)],说明林分的直径结构很难用同一分布族来描述。迫使人们寻找更好的方法,其中百分位方法[6]就是一例。但这种方法在方程选型上理论依据不足,并且效果也并不特别理想[2)]。是否有更好的描述直径结构的方法呢? 答案是肯定的。本文再引入一种非参数方法(Nonparametric estimate)[7]。这是比较新的一种方法,无需知道是否属于哪一分布族,就能对总体进行非常完美的描述。

1　非参数方法简介

它的基本思想是:

如果 $X_1, X_2, \cdots, X_n$ 是概率密度为 $f(x)$ 的总体的样本,则

$$f_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \tag{1}$$

就是 $f(x)$ 一个非参数核密度估计,它满足:

$$①\ f_n(x) \geq 0; \qquad ②\ \int f_n(x)\,dx = 1$$

其中,$K(x)$ 为核函数,一般取为适当的概率密度函数,h_n 为窗宽,它与样本容量有关,且 $h_n \rightarrow$

1995—12—12 收稿。

王雪峰助理研究员,唐守正(中国林业科学研究院资源信息研究所　北京　100091)。

* 本文属 1992 年国家自然科学基金项目"我国主要人工用材林生长模型、经营模型及优化控制"部分内容之一。

1) 李树生. 兴安落叶松天然林生长收获预测模型. 硕士论文, 1990.

2) 岳德鹏. 联立方程组在直径分布中的应用. 硕士论文, 1994.

0, $n\to\infty$时,$f_n(x)\to f(x)$ $(a.s.)$。可看出,它不需知道林分是否属于哪一分布族,就可对林分结构进行描述;同时它的假设很少,对于任给的两个$x_1<x_2$值,它都能够回答出在$[x_1, x_2]$内的株数或频率。很明显,用这种方法来描述直径结构具有得天独厚的优点。

在描述一个样本时,列点法最准确,但是仅能说明样本本身意义并不是很大,因为数理统计的核心是以样本估计总体;与此同时,列点法也没有消除随机因素的影响。那么在描述总体时究竟哪一种方法更好?本文准备采用计算机模拟技术,来探讨非参数核密度估计和直方图、列点法的优缺点。

理论上,$K(x)$为任意函数,但是从应用角度,$K(x)$多取为适当的概率密度函数。本文采用在统计上具有很多优良性质[7]的(2)式。

$$K(x)=\begin{cases}\dfrac{3}{4\sqrt{5}}\left(1-\dfrac{x^2}{5}\right) & |x|\le\sqrt{5}\\ 0 & |x|>\sqrt{5}\end{cases} \tag{2}$$

h_n的选取,是核估计的最关键问题,它直接影响着核估计精度。笔者通过大量的试验,认为把直方图的宽度与所选核函数最大值的乘积作为窗宽,能够满足应用要求,同时还能达到很高的精度。由于林业上大多以 2 cm 作径阶宽,故本文取$h_n=0.67$。

2 试验方法

2.1 比较原理

设一总体,直径累积分布为$F(x)$,从中抽取n个单元组成一个样本,其标志值为$x_1, x_2, \cdots, x_n$。本文采用柯尔莫哥洛夫检验法,比较列点法与核方法在描述总体时的优劣;而比较直方图与核方法时,采用卡方检验法。

2.1.1 列点法与核方法的比较

令
$$\lambda=\sup_x[|F_l(x)-F(x)|] \tag{3}$$

$$\lambda_k=\sup_x[|F_k(x)-F(x)|] \tag{4}$$

其中
$$F_l(x)=\frac{1}{n}\sum_{x_i\le x}n_i \tag{5}$$

$$F_k(x)=\int_0 f_n(y)\,dy=\frac{1}{nh_n}\int_0\sum_{i=1}^{n}K\left(\frac{y-x_i}{h_n}\right)dy=\frac{1}{nh_n}\sum_{i=1}^{n}\int_0 K\left(\frac{y-x_i}{h_n}\right)dy \tag{6}$$

$K(x)$取(2)式。因为λ_k和λ_l都近似遵从斯米尔诺夫—柯尔莫哥洛夫分布,所以,如果$\lambda_k<\lambda$,则说明在描述总体时,核方法要优于列点法。

2.1.2 直方图与核方法的比较

设样本分组数据为D_j, m_{hj} $(j=1, 2, \cdots, m)$,其中,D_j为j径阶中值,m_{hj}为j径阶株数。则核方法计算各径阶株数的计算式如下:

$$\begin{aligned}m_{ki}&=n\int_{i-2+D_0}^{i+D_0}f_n(y)\,dy=n\int_{i-2+D_0}^{i+D_0}\frac{1}{nh_n}\sum_{j=1}^{m}m_{h\cdot j}K\left(\frac{y-D_j}{h_n}\right)dy\\&=\frac{1}{h_n}\int_{i-2+D_0}^{i+D_0}\sum_{j=1}^{m}m_{h\cdot j}K\left(\frac{y-D_j}{h_n}\right)dy=\frac{1}{h_n}\sum_{j=1}^{m}m_{h\cdot j}\int_{i-2+D_0}^{i+D_0}K\left(\frac{y-D_j}{h_n}\right)dy\end{aligned} \tag{7}$$

D_0 为最小径阶下限, $K(x)$ 取(2) 式。下面我们取

$$X_h^2 = \sum_{j=1}^{m} \frac{(m_{hj} - np_j)^2}{np_j} \tag{8}$$

$$X_k^2 = \sum_{j=1}^{m} \frac{(m_{kj} - np_j)^2}{np_j} \tag{9}$$

式中, p_j 为直径落入第 j 径阶的概率。当 $X_k^2 < X_h^2$ 时, 则表明由样本描述总体时, 核方法要优于直方图方法。

2. 2　模拟总体

由于本试验对数据要求较高, 故采用计算机模拟方法, 生成遵从某一分布 $F(x)$ 的总体标志值, 从中随机抽取若干样本进行估计。

令 ξ 在(0, 1) 内为均匀分布, 若 $F^{-1}(x)$ 为 $F(x)$ 的反函数, 则 $\eta = F^{-1}(\xi)$ 的分布函数为 $F(x)$。例如, 对于 Weibull 分布

$$F(x) = 1 - e^{-\left(\frac{x-a}{b}\right)^c} \tag{10}$$

有:

$$\eta = a + b\left(\ln \frac{1}{1-\xi}\right)^{\frac{1}{c}} \tag{11}$$

这样, 只要随机产生(0, 1) 间的 ξ 值, 由(11) 式得到的 η 值即为遵从 Weibull 分布的直径值。由于 Weibull 分布均值为:

$$W = b\Gamma(1 + 1/c) + a \tag{12}$$

故欲产生平均直径为 D 的直径值, 可由上式计算。本试验产生 4 种不同平均直径 D 的总体数据, 总体单元数为 10 000 株。各总体单元参数见表 1。

表 **1**　Weibull 分布参数

D	a	b	c
10	3. 5	6. 50	1. 0
15	5. 3	9. 70	1. 0
20	7. 0	14. 65	1. 9
25	8. 8	18. 24	2. 6

2. 3　样本组织

在实际外业测定中, 样地面积不可能很大, 为了不至于使本试验失真, 将对抽取的单元数进行一定限制。

已知, $N = S_t(D/20)^{-\beta}$, 其中 S_t 为密度指数, N 为每公顷株数。对于长白落叶松(*Larix olgensis* Henry) , β 取 1. 68。采用下式抽取试验的样本单元数。

$$n = \frac{S_t}{10}\left(\frac{D}{20}\right)^{-1.68} \tag{13}$$

则, 这大约相当于样地面积为 0. 1 hm^2 的长白落叶松实测株数。S_t 分别取 400、600、800、1 000、1 200、1 400、1 600, 对于不同的 D, 由(13) 式将分别得到应抽取的样本单元数。这样, 共产生 4 种不同直径、7 种不同密度林分的模拟数据, 每种组合产生 10 个样本(重复) , 按 2. 1 节的原理进行计算、分析。

3　结果分析

3. 1　列点法与核方法的比较

由 2. 1. 1 节的(3) 、(4) 式, 得到柯尔莫哥洛夫检验值见表 2。

总的说, 共进行 280 次比较, 核方法优于列点法达 254 次, 而列点法优于核方法的次数仅

26 次; 还可以看出, 参加试验的样本单元数越少, 核方法描述总体的相对效果越好。

表 **2** 核方法与列点法的柯尔莫哥洛夫检验比较

密度指数	直径 d= 10 cm			直径 d= 15 cm			直径 d= 20 cm			直径 d= 25 cm		
	n	B	C	n	B	C	n	B	C	n	B	C
400	128	10	9	64	10	10	40	10	10	27	10	10
600	192	10	7	97	10	10	60	10	10	41	10	10
800	256	10	7	129	10	10	80	10	10	54	10	10
1 000	320	10	6	162	10	10	100	10	10	68	10	10
1 200	384	10	6	194	10	8	120	10	10	82	10	10
1 400	448	10	6	226	10	10	140	10	10	96	10	10
1 600	512	10	5	259	10	10	160	10	10	109	10	10
总 计		70	46		70	68		70	70		70	70

注: n 为抽取的样本单元数, B 为重复次数, C 为核方法 λ_k 小于列点法 λ_l 的次数; $x \in [0, 70]$, 步长 0.05。

3.2 核方法与直方图方法的比较

首先, 从总体中随机抽取样本, 抽取的样本单元数按(13)式进行计算; 然后, 对数据以 2 cm 径阶宽度进行分组, 得到 m_{hj}; 再由 2.1.2 节中的(7)式计算 m_{ki}。然后由(8)、(9)式计算核方法与直方图方法的卡方值得到表 3。

表 **3** 分组数据的核方法与直方图的卡方检验比较

密度指数	直径 d= 10 cm			直径 d= 15 cm			直径 d= 20 cm			直径 d= 25 cm		
	n	B	C	n	B	C	n	B	C	n	B	C
400	128	10	9	64	10	10	40	10	10	27	10	10
600	192	10	8	97	10	10	60	10	10	41	10	9
800	256	10	6	129	10	10	80	10	9	54	10	10
1 000	320	10	6	162	10	10	100	10	9	68	10	9
1 200	384	10	7	194	10	10	120	10	10	82	10	8
1 400	448	10	5	226	10	10	140	10	10	96	10	9
1 600	512	10	5	259	10	10	160	10	8	109	10	9
总 计		70	46		70	70		70	66		70	64

注: n 为抽取的样本单元数, B 为重复次数, C 为核方值核方法小于直方图方法的次数。

从表 3 的结果看, 进行 280 次试验, 核方法优于直方图的次数为 246 次, 直方图优于核方法的次数为 34 次; 另外, 从总的情况看, 也是参加试验的样本单元数越少, 核方法描述总体的相对效果越好。

3.3 多峰总体核方法与列点法、直方图的效果分析

由于 3.1 节和 3.2 节的模拟数据都来自单峰总体, 为进一步比较多峰总体情况, 将平均直径分别为 10 cm 和 25 cm 的两个 Weibill 分布函数叠加而成一个新的总体。其中, 平均直径 10 cm 的 3 000 株, 平均直径为 25 cm 的 7 000 株。从总体中抽取的样本数由(13)式得到。按上面的方法, 在比较核方法与列点法时用柯尔莫哥洛夫检验; 而比较核方法与直方图时采用卡方检验法(见表 4)。

可以看出, 对于多峰总体, 结论同 3.1 节和 3.2 节是一样的, 仍然是核方法模拟总体的效

果要比另外两种方 法要好。为进一步从统计角度来检验以上结论的正确性, 对以上结果进行了符号检验, 结论见表 5、6。

由表 5、6 可知, 不论是单峰还是多峰总体, 符号检验结果都表明核方法既优于列点法, 又优于直方图; 特别是当总体为多峰时, 用核方法模拟总体要远远优于列点法及直方图法。

表 **4**　多峰总体核方法与列点法、直方图的效果分析

密度指数	核方法与列点法			核方法与直方图		
	n	*B*	*C*	*n*	*B*	*D*
400	128	10	10	128	10	10
600	192	10	10	192	10	10
800	256	10	9	256	10	10
1 000	320	10	10	320	10	9
1 200	384	10	10	384	10	9
1 400	448	10	9	448	10	9
1 600	512	10	10	512	10	9
总　计		70	68		70	66

注: *C* 为柯尔莫哥洛夫检验中核方法 λ_k 小于列点法 λ_l 的次数; $x \in [0, 70]$, 步长 0.05;
D 为卡方检验中卡方值核方法小于直方图方法的次数;
n 为抽取的样本单元数, *B* 为重复次数。

表 **5**　单峰总体符号检验结果(显著水平 0.95)

平均直径	核方法与列点法		核方法与直方图		临界值
	s	*n*	*s*	*n*	s_a
10	46	70	46	70	44
15	68	70	70	70	44
20	70	70	66	70	44
25	70	70	64	70	44

注: *s* 为出现 "+" 的次数; *n* 为试验次数。

表 **6**　多峰总体符号检验结果(显著水平 0.95)

核方法与列点法		核方法与直方图		临界值
s	*n*	*s*	*n*	s_a
68	70	66	70	44

4　结论

核方法是古老的直方图方法的自然发展, 理论上它具有直方图方法无法比拟的优点, 通过本文的研究, 从模拟角度证明了核方法要比直方图方法好。

实际上, 核方法是一种修匀方法, 它消除了随机误差的影响, 故用它模拟总体时要比直接用列点法效果好。

对于多峰总体, 与直方图和列点法相比, 核方法的优势更突出。

当样本数很少时, 核方法模拟总体的效果要远远优于列点法及直方图法; 随着样本数增加, 核方法的优点渐减, 最后同直方图和列点法趋于一致。这与实际问题是相符合的。因为样本数很大时, 无论用哪一种方法, 都与总体接近。此时, 方法本身将变得不重要了, 那么, 在这种时候, 应该选择最简单的方法。

参　考　文　献

1　Bliss C L, Reinker K A. A lognormal approach to diameter distribution in even-aged stands. For. Sci., 1964, (10): 350～360.

2　Nelson T C. Diameter distribution and growth of loblolly pine. For. Sci., 1964, (10): 105～115.

3　Clutler J L, Bennett F A. Diameter distributions in old-field slash pine plantations: Ga. Forest Res. Counc. Rcp., 1965, 13: 9.

4　Hafley W L, Schreuder H T. Statistical distributions for fitting diameter and height data in even-aged stands. Can. J. For. Res., 1977, (7): 481～487.

5　Bailey R L, Dell T R. Quantitying diameter distributions with the Weibull function. For. Sci., 1973, (19): 97～104.

6　Borders B E, Souter R A, Bailey R L, et al. Perntile-baded distributions characterize forest stand stables. For Sci., 1987, (33): 570～576.

7　陈希儒, 方兆本, 李国英, 等. 非参数统计. 上海: 科学技术出版社, 1989.

Simulating Diameter Structure: A Comparison of Nonparametric Kernel Method, Histogram Method and Diameter List Method

Wang Xuefeng *Tang Shouzheng*

Abstract The population diameter structure can well be described by using nonparametric kernel method based on the sample data without assuming the population diameter distribution. The results of the computer simulation have showed that nonparametric kernel method is better than both histogram method and diameter list method in obtaining the population diameter structure.

Key words simulation, nonparametric kernel density estimation, diameter distribution

Wang Xuefeng, Assistant Professor, Tang Shouzheng (The Research Institute of Forest Resource Information Techniques, CAF Beijing 100091).

《桉树营养》评介

由 P M Attiwill 和 M A Adams 主编和 10 多个国家 30 多位研究桉树的学者参加编写的《桉树营养》(Nutrition of Eucalyptus)一书,最近由澳大利亚联邦科工组织(CSIRO)出版。全书 448 页,49 幅彩图,精装,售价 150 澳元(约合人民币 1000 元)。

此书概括了所有关于桉树营养的研究成果和大量文献,当前桉树人工林营养管理的进展和营养缺乏的诊断。澳大利亚学者论述了澳大利亚森林土壤中的磷、土壤对桉树进化的影响,澳大利亚自然景观中的桉树分布,桉树的营养生理和养分循环及林分经营。澳大利亚、新西兰、南非、阿根廷、巴西、智利、葡萄牙、中国和印度等国学者分别撰述了各自国家桉树人工林的培育和施肥。该书最后一章是桉树营养缺乏诊断。此书涉及 110 种桉树,引用 245 篇有关桉树研究文献,是研究和经营桉树人工林不可多得或缺的重要参考书。如需购买者,可与中国林科院林研所王豁然联系。

(王豁然)