基于最大熵原理的浙江毛竹胸径分布及测量不确定度评定

刘恩斌*, 周国模, 葛宏立
(浙江林学院环境科技学院, 临安 311300)

摘要:应用最大熵原理构造了测树因子概率分布的统一模型,这样构造的模型具有明确的解析表达式,并能克服常用方法无法解释测树因子服从某种概率分布的真正原因,从而为测树因子统计分布建模提供了一种有效方法。使用 1-3 阶样本矩、1-4 阶样本矩与 1-5 阶样本矩,用所构建的概率分布统一模型分别对浙江省域毛竹胸径分布分别作了仿真试验,结果表明当采用 1-4 阶样本矩时,仿真效果最好,而且比通过假设检验的 Weibull 分布仿真结果理想:(1)图形非常相似,对实测数据都能很好的模拟;(2)最大熵法的离差平方和为 0.00018, Weibull 分布的为 0.00045[1]。由于各种系统与非系统的原因,都会影响测量结果的准确性,对所构建的模型作了不确定度评定,表明结果具有很大的可靠性,测量结果的估计:7.85100,测量结果的标准不确定度:1.82710,置信概率:0.96020。

关键词:最大熵原理;概率分布模型;Weibull 分布;测量不确定度
文章编号:1000-0933(2009)01-0086-06 **中图分类号**:Q141, Q948 **文献标识码**:A

Examination of moso bamboo's diameter probability distribution and evaluation of measurement uncertainty with the maximum entropy theory

LIU En-Bin*, ZHOU Guo-MO, GE Hong-Li
School of Environmental Technilogy, Zhejiang Forestry University, Linan 311300, *China*

***Acta Ecologica Sinica*, 2009, 29(1): 0086 ~ 0091.**

Abstract: This paper used a maximum entropy theory to establish a general probability distribution model for tree measurement parameters. This model has an explicit explanatory expression and overcomes some problems occurred in the traditional methods that the reasons why the tree measurement parameters obeyed certain probability distribution cannot effectively be explained. Therefore, this model provides a new way to establish the statistical distribution for tree measurement parameters. The established general probability distribution model was used to simulate the Moso bamboo's diameter distribution in Zhejiang province based on 1-3, 1-4, and 1-5 stage sample moments, respectively. The results indicated that using 1-4 stage sample moment provided the best simulation performance, and provided even better effects than that using Weibull distribution. Both maximum entropy theory and Weibull distribution have similar features that can effectively simulate the reference data. The sum of square deviation is 0.00018 based on maximum entropy theory and 0.00045 based on Weibull distribution. Because different system and non-system factors can affect the reliability of the estimates, the established models was used to evaluate the measurement uncertainty, indicating the reliable results with estimates of 7.85100, standard uncertainty of 1.8271, and confidence probability of 0.9602.

Key Words: maximum entropy theory; probability distribution model; Weibull distribution; measurement uncertainty

---

基金项目:国家自然基金资助项目(30771725);国家自然基金资助项目(30700638)
收稿日期:2008-04-14; **修订日期**:2008-09-28

* 通讯作者 Corresponding author. E-mail: liuenbin2001@yahoo.com.cn

在林业与生态实践中,经常要用到测树因子统计分布模型,惯用的方法是:把测树因子的调查数据套到一个适合的分布上去,然后检验,若分布通过检验则认为该分布是指导工作的理论模型[1,2]。由于惯用方法无法回答概率分布是凑巧吻合了这些数据,还是产生了这些数据的母体,故该方法具有的缺陷是在给定的显著水平下,当随机变量对多种概率分布均不拒绝时,无法确定那种概率模型更能准确描述测树因子的分布规律,也就是无法解释测树因子服从某种概率分布的真正原因。因此如何找到一个通用的生成概率密度函数的方法对深刻了解测树因子分布规律的本质具有更广泛的实践意义。

要构造一个通用的概率密度函数,必须有一个构造的标准,那就是使所构造的模型,在数据不充分的情况,既要与已知的数据相吻合又必须对未知的部分作最少的假定,即对数据的外推或内插采取最超然的态度。因此模型的构造过程可以认为是从数据中提取信息的过程,而信息来自两个部分:一是已知数据,二是由于数据不完全而不得不对未知部分所作的假定,这种假定相当于人为地"添加"了一些信息。由于熵是信息论中的一个基本概念,是用以度量信息源不确定性的量,所以熵可以用来度量测树因子的不确定性,熵最大就意味着获得的总信息量最少,即所添加的信息最少,所以最大熵是超然的。因此对于只有测量数据样本的情况,若没有充足的理由来选择某种解析分布函数时,可通过最大熵方法来确定出最不带倾向性的总体分布形式。本文以最大熵原理为基础,给出了由测量样本确定测树因子概率分布的方法,将该方法应用于浙江毛竹的胸径分布,并对该分布测量结果的估计及测量不确定度进行了评定。

1　最大熵原理[3,4]

利用样本信息的一种简便方法是计算样本的各阶矩。下面对连续型随机变量的最大熵方法作一详细阐述,对于离散随机变量可以做相应的推导:

$$S = -\int_R f(x)\ln[f(x)]\,\mathrm{d}x = \text{最大值} \tag{1}$$

$$\int_R f(x)\,\mathrm{d}x = 1 \tag{2}$$

$$\int_R x^i f(x)\,\mathrm{d}x = m_i, i = 1,2,\cdots,m \tag{3}$$

式中,S 为信息熵,$f(x)$ 为测树因子的概率密度函数,m 为所用矩的阶数,m_i 为第 i 阶原点矩,其值可用样本确定,R 为积分区间。

下面通过调整 $f(x)$,使其熵达到最大,设 $\bar{S}$ 为拉格朗日函数,$\lambda_0,\lambda_1,\cdots,\lambda_m$ 为拉格朗日乘子。则:

$$\bar{S} = S + (\lambda_0 + 1)\left[\int_R f(x)\,\mathrm{d}x - 1\right] + \sum_{i=1}^{m}\lambda_i\left(\int_R x^i f(x)\,\mathrm{d}x - m_i\right) \tag{4}$$

令 $\mathrm{d}\bar{S}/df(x)$ 等于零,得:

$$-\int_R\{\ln f(x) + 1\}\,\mathrm{d}x] - (\lambda_0 + 1)\int_R \mathrm{d}x - \sum_{i=1}^{m}\lambda_i\left(\int_R x^i\mathrm{d}x\right) = 0 \tag{5}$$

从(5)式可以解得:

$$f(x) = \exp\left(\lambda_0 + \sum_{i=1}^{m}\lambda_i x^i\right) \tag{6}$$

(6)式就是最大熵原理推出的测树因子概率密度函数解析式。

2　算法实现[5,7]

前面提到样本的各阶矩能反映样本的信息,所以本文用样本矩实现最大熵算法。

把(6)式代入(2)式得:

$$\int_R \exp\left(\lambda_0 + \sum_{i=1}^{m}\lambda_i x^i\right)\mathrm{d}x = 1 \tag{7}$$

整理后得:

$$\lambda_0 = -\ln\left(\int_R \exp\left(\sum_{i=1}^{m}\lambda_i x^i\right)\mathrm{d}x\right) \tag{8}$$

将(8)对 λ_i 求微分得:

$$\frac{\partial \lambda_0}{\partial \lambda_i} = -\frac{\int_R x^i \exp\left(\sum_{i=1}^{m} \lambda_i x^i\right) dx}{\int_R \exp\left(\sum_{i=1}^{m} \lambda_i x^i\right) dx} \tag{9}$$

$$m_i = \frac{\int_R x^i \exp\left(\sum_{i=1}^{m} \lambda_i x^i\right) dx}{\int_R \exp\left(\sum_{i=1}^{m} \lambda_i x^i\right) dx} \tag{10}$$

通过(10)式可以建立 $\lambda_1, \lambda_2, \cdots, \lambda_m$ 的 m 个方程组,求出这 m 个参数后,代入(8)式求出 λ_0,为了便于数值计算,将(10)式改为:

$$1 - \frac{\int_R x^i \exp\left(\sum_{i=1}^{m} \lambda_i x^i\right) dx}{m_i \int_R \exp\left(\sum_{i=1}^{m} \lambda_i x^i\right) dx} = r_i \tag{11}$$

式中,r_i 为残差,用 Matlab 的 lsqnonlin 函数进行优化求解。由测量数据样本,用最大熵方法确定概率分布算法的步骤如下:

(1)求样本各阶矩并确定积分上下界。对于大多数分布来说,用 3 ~ 6 阶矩就可以作出比较理想的密度函数。积分上下界可由样本的分散范围来确定。

(2)确定 $\lambda_1, \lambda_2, \cdots, \lambda_m$ 的初值。由于 Matlab 的 lsqnonlin 函数对初值选取比较敏感,如果初值选的不好,迭代就不收敛,因此在此提出如下的初值选取方法:先对(6)取自然对数做线性化处理,然后用普通最小二乘拟合参数,得到的结果作为 $\lambda_1, \lambda_2, \cdots, \lambda_m$ 的初值。

(3)按(11)建立残差表达式。

(4)把 $\lambda_1, \lambda_2, \cdots, \lambda_m$ 的初值代入 lsqnonlin 函数进行优化求解。

(5)把求得的 $\lambda_1, \lambda_2, \cdots, \lambda_m$ 代入(8)式求出 λ_0。

从最大熵的基本原理可以看出,该方法仿真结果的精度与所取样本矩的阶数有关,而样本的矩又与样本的容量有关。

3 仿真案例

数据材料与 Weibull 分布的假设检验见参考文献[1],本文分别采用 1-3、1-4、1-5 阶样本矩,用如上算法拟合参数见表 1。

表 1 采用不同样本矩所得模型参数与评价指标对比

Table 1 Contrasting Model's paramater and its evaluation index using different sample moment

样本矩 Sample moment	λ_0	λ_1	λ_2	λ_3	λ_4	λ_5	离差平方和 Minimum variance	R^2
1-3	-5.95901	0.83967	-0.01077	-0.00343	0	0	0.00033	0.99683
1-4	2.37831	-3.16391	0.67158	-0.05251	0.00126	0	0.00018	0.99906
1-5	-11.9455	5.1340	-1.1740	0.14500	-0.00890	0.00020	0.00025	0.99782

当采用 1-3 阶样本矩时,最大熵方法构造模型与 Weibull 分布的模拟结果见图 1。

当采用 1-4 阶样本矩时,最大熵方法构造模型与 Weibull 分布的模拟结果见图 2。

当采用 1-5 阶样本矩时,最大熵方法构造模型与 Weibull 分布的模拟结果见图 3。

从图 1 ~ 图 3 可以看出,当采用 1-3 阶样本矩与 1-5 阶样本矩时,最大熵方法与 Weibull 分布对浙江省毛竹胸径分布模拟结果相差不大。这主要是由于最大熵构造的模型可以逼近 Weibull 概率密度函数,事实上可

对最大熵函数做如下推导:

$$\left(a_0 + \sum_{i=1}^{m} a_i x^i\right)\exp\left(c_0 + \sum_{i=1}^{m} c_i x^i\right) = \exp\left(\frac{1}{e}\left(a_0 + \sum_{i=1}^{m} a_i x^i\right)\right)\exp\left(c_0 + \sum_{i=1}^{m} c_i x^i\right)$$

$$= \exp\left(\lambda_0 + \sum_{i=1}^{m} \lambda_i x^i\right)$$

从这里可以看出最大熵模型由 $a_0 + \sum_{i=1}^{m} a_i x^i$ 与 $\exp\left(c_0 + \sum_{i=1}^{m} c_i x^i\right)$ 组成,而前者与后者的幂都是一维连续

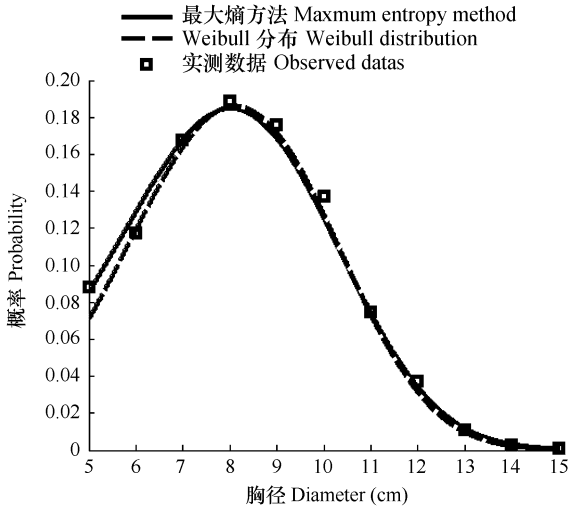

图 1　最大熵法构造分布函数曲线、Weibull 分布曲线与实测数据对比

Fig. 1　Fitted curve of maximum entropy, curve of Weibull and scatter plot of observed datas

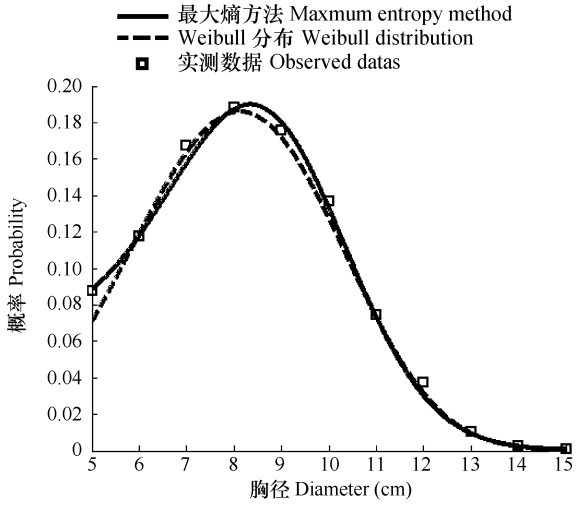

图 2　最大熵法构造分布函数曲线、Weibull 分布曲线与实测数据对比

Fig. 2　Fitted curve of maximum entropy, curve of Weibull and scatter plot of observed datas

函数空间基$\{1, x, x^2, \cdots, x^m, \cdots\}$(当取至 m 时达到精度要求)的线性组合,而 Weibull 概率密度函数的$\frac{c}{x-a}\left(\frac{x-a}{b}\right)^c$与$e^{-\left(\frac{x-a}{b}\right)^c}$则分别是最大熵 2 组成部分的其中一个元素,因此 Weibull 概率密度函数可用 $\exp\left(\lambda_0 + \sum_{i=1}^{m} \lambda_i x^i\right)$来逼近;当采用 1-4 阶样本矩时,模拟效果最好,此时最大熵构造的模型与实测数据几乎完全吻合,仿真结果要比 Weibull 分布好,主要是由于浙江省毛竹胸径分布符合 Weibull 分布是建立在事先假定的基础上,包含了人为主观的因素,而最大熵方法采用最超然的态度,从而使主观"添加"的信息最小。为什么样本矩取到 4 阶仿真效果最好呢? 主要是由于取的样本矩太小,所损失的信息太多,取的太大会影响模型的结构[8],或者是由于采用 1-5 阶样本矩时,本文所用的算法没有找到全局最优解,可采用其他优化算法或几种优化算法的组合进行搜索。用最大熵法得出的理论值与实测值存在差异的原因是样本的矩与总体理论的

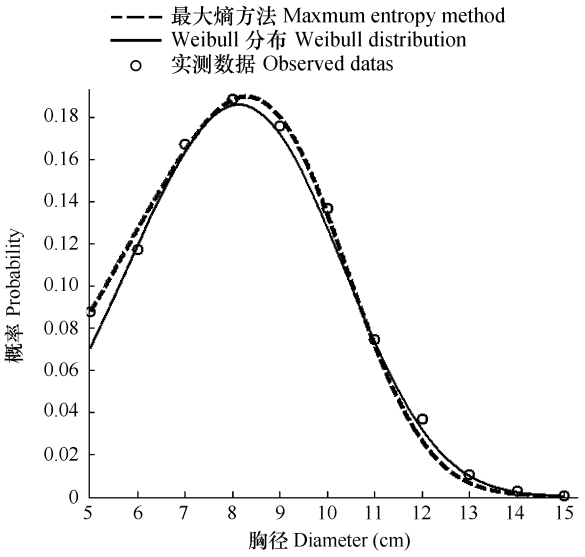

图 3　最大熵法构造分布函数曲线、Weibull 分布曲线与实测数据对比

Fig. 3　fitted curve of maximum entropy, curve of Weibull and scatter plot of observed datas

矩存在一定的偏差,故求样本矩时,应采用较多的样本。关于最大熵方法与其它分布的对比研究见参考文献[3,6]。从文献[3,6]与本文的仿真模拟说明最大熵法确实是用测量样本推算其概率分布的通用方法。

在林业与生态实践中,由于测量方法、测量程序、测量系统、测量人员的技术、数据处理使用的参数与方法等因素造成测量误差是不可避免的,这样就会影响测量的准确性,因此必须要对测量结果进行不确定度评定。

4 测量结果的不确定度评定

在本文中,当取 1-4 阶样本矩,用最大熵法求得测量样本的概率密度函数 $\hat{f}(x)$ 后,就可以对测量结果的不确定度进行评定[6,9]。

测量结果的估计为:

$$\hat{x} = \int_a^b x\hat{f}(x)\,\mathrm{d}x \tag{12}$$

测量结果的标准不确定度为:

$$u = \sqrt{\int_a^b (x-\hat{x})^2\hat{f}(x)\,\mathrm{d}x} \tag{13}$$

由于本文所拟合的曲线是非对称的,所以置信概率可用下式表示:

$$p = \int_{\hat{x}-(1+a)ku}^{\hat{x}+(1-a)ku} \hat{f}(x)\,\mathrm{d}x \tag{14}$$

式中,k 为包含因子,一般取值为 2 ~ 3,a 为不对称系数,可由偏度与峰度来确定[3,4]。

把最大熵法构造的概率分布函数代入(12)、(13),求得仿真案例测量结果的估计:7.85100,测量结果的标准不确定度:1.82710。同样,拟合得到的 Weibull 分布的值分别为:7.72880 与 1.84270。从这 2 组数值也可以看出,用最大熵法推出的概率分布其性能要优于 Weibull 分布。

把最大熵法构造的毛竹概率分布函数代如(14),则当 $k = 2$ 时,置信概率 $P = 0.96020$,是一个很大的概率,说明本文用最大熵构造的模型可信度很高。

5 结论

从最大熵推导概率分布的过程可以看出,对于不同的约束条件,可以得到不同的概率分布函数,这表明最大熵方法可以作为众多概率分布的统一理论基础,而且用最大熵产生的概率密度函数属于一种理论模型,它除了具有明确的数学解析式外,还具有计算简单的特点。综合前面的推导与仿真,可以得到如下结论:

(1)最大熵方法推导出的测树因子概率分布函数具有如下的特点:第一,统一的理论基础,明确的函数解析式,在事先没有充分的理由确定测树因子到底服从什么样的概率分布时,就用最大熵方法来构造测树因子的概率分布函数;第二,可以解释说明测树因子为什么符合某个概率分布的真正原因,即既要最大限度的利用现有数据又要使主观“添加”的信息最小;第三,其模拟精度取决于样本容量、初值的选取;第四,能最大限度的利用现有数据所提供的信息,所以不需要再做假设检验;第五,与正态分布、指数分布、伽马分布、瑞利分布及 Weibull 分布等有一个共同的特点,即都含有测量因子的指数函数,但最大熵方法构造的模型其指数幂是多项式,因此本文所构建的模型可以逼近任一常用概率分布函数,可作为胸径分布的统一模型。

(2)本文对最大熵方法所构造的样本总体概率分布函数的测量结果进行了评定,表明结果是可靠的。

References:

[1] Zhou G M, Liu E B, Liu A X, Zhou Y F. The algorithm update of Weibull Distribution parametric identification and its application on measuring the distribution of diameter and age of Moso bamboo forests in Zhejiang Province. Acta Ecologica Sinica, 2006, 26(9): 2918 - 2924.

[2] Hong L X, Du G J, Zhang Q R. Weibull D. B. H. distributions and their dynamic predictions in the natural uneven-aged evergreen broad-leaf juvenile forest. Acta Phytoecologica Sinica, 1995, 19(1): 29 - 42.

[3] Siddall, J. N. engineering probability design: principle and application. Beijing: Beijing Science Press, 1989.
[4] Wu N L, Yuan S Y. The maximum entropy method. Changsha: Hunan Science & Technic Press, 1991.
[5] Diao Y F, Wang B D, Liu J. Study on distribution of flood forecasting errors by the method based on maximum entropy. Journal of Hydraulic Engineering, 2007, 38(5): 591 - 595.
[6] Zhu J M, Guo B J, Wang Z Y, Xia X T. Study on Evaluation of Measurement Result and Uncertainty Based on Maximum Entropy Method. Electrical Measurement & Instrumentation, 2005, 42(476): 5 - 8.
[7] Jon Lee. Constrained maximum entropy sampling. Operations Research, 1998, 46(5): 655 - 664.
[8] Lin cheng sen. Numerical Analysis. Beijing: Science Press, 2006.
[9] Qian S S. Measurement uncertainty. Beijing: Tsinghua University Press, 2002.

参考文献:

[1] 周国模, 刘恩斌, 刘安兴, 周宇峰. Weibull 分布参数辨识改进及对浙江毛竹林胸径年龄分布的测度. 生态学报, 2006, 26(9): 2918 ~ 2924.
[2] 洪利兴, 杜国坚, 张庆荣. 天然常绿阔叶异龄幼林胸径的 Weibull 分布及动态预测. 植物生态学报, 1995, 19(1): 29 ~ 42.
[3] 希德尔 J. N. 工程概率设计: 原理和应用. 北京: 北京科学出版社, 1989.
[4] 吴乃龙, 袁素云. 最大熵方法. 长沙: 湖南科学技术出版社, 1991.
[5] 刁艳芳, 王本德, 刘冀. 基于最大熵原理方法的洪水预报误差分布研究. 水利学报, 2007, 38(5): 591 ~ 595.
[6] 朱坚民, 郭冰菁, 王中, 夏新涛. 基于最大熵方法的测量结果估计及测量不确定度评定. 电测与仪表, 2005, 42(476): 5 ~ 8.
[8] 林成森. 数值分析. 北京: 科学出版社, 2006.
[9] 钱绍圣. 测量不确定度. 北京: 清华大学出版社, 2002.