

基于纵向数据非线性混合模型的杉木林 优势木平均高研究

李春明

(中国林业科学研究院资源信息研究所,北京 100091)

摘要:以江西省大岗山实验局不同初植密度的杉木林为研究对象,选择修改的 Richards 模型形式,考虑样地效应,采用 SAS 软件进行非线性混合效应模型的模拟,利用 AIC 和 BIC 值评价模型模拟效果。在此基础上考虑优势木平均高连续观测数据的时间序列相关性,并把初植密度以哑变量形式考虑进去,再进行混合模型的模拟。最后,利用验证数据对混合模型方法与传统的非线性回归模拟方法进行精度比较。研究结果表明,修改的 Richards 形式的优势木平均高与林龄关系的非线性混合效应模型,其估计精度比传统的回归模型估计精度明显提高,增加随机效应参数个数能够提高模型的估计精度。一阶自回归误差结构矩阵模型在解释优势木平均高的时间序列相关性时不仅提高了混合模型的模拟精度,而且能够很好的表达连续观测数据间误差分布情况;同时考虑样地的随机效应、观测数据的时间序列相关性,不同初植密度的混合模型模拟精度比传统的非线性回归方法模拟精度高。

关键词:杉木;优势木平均高;非线性混合效应模型;时间序列相关性

中图分类号:S791.27

文献标识码:A

Study on Dominant Height Growth of Fir Plantations Based on a Nonlinear Mixed Modeling Approach for Longitudinal Data

LI Chun-ming

(Research Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China)

Abstract: The improvement on the dominant height growth implies in better productivity estimation due to the forest height growth is directly related with the site characteristics and forest productivity. A modified Richards growth model with nonlinear mixed effects is simulated used SAS software for modeling fir plantation dominant height growth in conjunction with different plantation density in Dagangshan Experiment Bureau of Jiangxi Province. Akaike Information Criterion(AIC) and Bayesian Information Criterion(BIC) were used in model performance evaluation. Within-plot time series error autocorrelation of dominant height growth data and different plantation density expressed with dummy variable were taken into account in mixed model. Finally, the precision of mixed models was compared with the precision of conventional nonlinear ordinary regression analysis method based on validation data. The result showed that the precision of modified Richards forms nonlinear mixed effect model which takes into account plot's random effect was better than that of conventional regression model. Increasing the number of random effect parameter can increase the precision of model. First-order autoregressive error model in explaining time series error autocorrelation of dominant height growth not only improved simulated precision, but also can describe error distribution of sequence observation data; The precision of mixed model considering plot random effects, time series error autocorrelation and different plantation density at one time is better than that of ordinary regression

收稿日期:2010-03-20

基金项目:中央级公益性科研院所基本科研业务费专项资金课题(IFRIT200901)及国家“十一五”科技支撑课题“东北天然林保护与可持续经营技术试验示范”(2006BAD03A08)部分研究内容。

作者简介:李春明(1975—),男,河北承德人,助理研究员,在读博士,主要从事森林生长模型方面的研究。

analysis method.

Key words: fir; dominant height growth; nonlinear mixed effect model; time series error autocorrelation

在评价森林生产力时,通常用地位级和地位指数来表示。由于林分平均高(反映地位级)受经营措施的影响较大,尤其在采伐的前后,立地质量没有任何变化,但林分的平均高却有很大的变化,表现出不稳定性,采用林分的优势木平均高(反映地位指数)就能够避免这一现象发生^[1]。因此在进行不同立地条件下林分的生长对比时,需要对优势木平均高准确估计。为达到此目的,常常需要对多个样地不同时间内的树高进行多次重复观测,这些多次测量的数据称其为纵向数据。这些数据存在着共性,即重复测量是相关的,且相邻数据相关性很大,而离得较远的数据随着时间间隔增大,相关性减小。另外,这些数据之间不满足独立性要求,而且存在着异方差性。而传统的统计分析方法要求数据满足独立性和常量方差,且很少能够考虑自相关性。因此利用混合效应模型方法对纵向数据进行研究可以得到更加符合现实的结果。不仅能够提高优势木平均高的估测精度,而且能够反映不同样地或立地条件下森林生产力的差异。

目前很多学者采用混合模型的方法对林分优势木平均高进行预测。Hall 和 Bailey^[2]以美国乔治亚州的火炬松(*Pinus taeda* L.)为研究对象,利用多水平非线性混合模型方程组构造了 Chapman-Richards 形式的优势高曲线。Fang 和 Bailey^[3]利用修改的三参数 Richards 生长模型构造了不同经营方式下美国乔治亚州和佛罗里达州湿地松(*Pinus elliottii* Engelm.)的优势木平均高的混合效应模型。Calegario 等^[4]考虑样地的随机效应,为采用不同无性繁殖技术处理的杂交桉树(*Eucalyptus* sp.)构造了 Logistic 形式的优势高和林龄关系的非线性混合模型。李永慈^[5]以杉木(*Cunninghamia lanceolata* (Lamb.) Hook.)优势高生长数据为研究对象,建立了对数 Schumacher 混合生长模型和 Logistic 混合生长模型。这些学者的研究结论认为与传统的最小二乘法相比,混合模型方法不仅提高了优势木平均高的估计精度,而且还可以反映不同林分的总体趋势和差异。

对许多树种来说,林分优势木平均高的生长不受林分密度影响。近来的研究发现,杉木在同一地区、相同的立地条件下由于初植密度不同,其优势高生长过程有较大差异^[6-7]。为了提高优势木平均高的估计精度及评价初植密度的影响,本文以江西

省大岗山实验局不同初植密度的人工杉木林为研究对象,考虑样地间的随机效应和观测的时间序列相关性,把初植密度作为哑变量的形式考虑到模型中去,选择变化后的 Richards 模型形式,通过变换混合效应参数个数来构造优势木平均高和林龄关系的非线性混合效应模型,并与传统最小二乘模拟方法进行精度比较。

1 数据

选择了中国林科院林业研究所设置在江西省大岗山实验局不同初植密度的人工杉木林进行研究。杉木数据为 5 个区组,共计 16 块杉木样地,样地大小为 0.06 hm²。初植密度分别为大致每公顷 1 667、3 333、5 000、6 667 和 10 000 株,分别在林龄为 5、6、7、8、9、10、11、12、14、16 年时测量了样地内每木胸径和树高。选择 6 株优势木和亚优势木的树高进行平均计算,得到林分的优势木平均高。具体数据见表 1。

表 1 人工杉木林样地概况

区组	样地号	初植密度/ (株·hm ⁻²)	优势木平均高/m	胸径/cm
I 区	1	1 667	4.1~15.9	3.8~16.9
	2	1 667	3.8~12.8	3.2~14.6
	3	1 667	4.1~15.4	3.7~17.0
II 区	4	3 317	4.1~13.9	3.7~13.5
	5	3 267	4.0~12.1	3.0~11.7
	6	3 283	3.8~14.6	3.0~13.4
III 区	7	5 000	4.2~13.1	3.4~11.0
	8	4 983	3.9~12.1	2.5~10.4
	9	5 000	3.9~12.9	2.7~10.8
IV 区	10	6 583	4.0~11.2	2.9~9.3
	11	6 567	3.6~11.1	2.3~9.4
	12	6 483	3.8~11.4	2.5~9.6
V 区	13	9 983	4.2~11.6	2.9~8.5
	14	9 600	4.0~11.1	2.7~8.3
	15	9 417	3.7~12.5	2.2~9.1
	16	9 767	4.0~12.6	2.7~8.8

2 方法

2.1 Richards 基础模型

在估测林分优势木平均高时,通常用到 3 参数的 Richards 形式^[8],具体如(1)式:

$$HD = \beta_1 \{1 - \exp(-\beta_2 t)\}^{\beta_3} + \varepsilon \quad (1)$$

其中, HD 指优势木平均高, β_1 为优势高渐近参

数, β_2 为尺度参数, β_3 为形状参数, t 为林龄, ε 为随机误差。

Fang 和 Bailey^[3] 认为(1)式的 β_1 是最不稳定的参数, 所以为了避免优势高和地位指数之间的不相容性, 对(1)式进行了改进。本文即采用 Fang 和 Bailey 提出的方程, 具体形式如(2)式:

$$HD = \beta_1 \left(\frac{1 - e^{-\beta_2 t}}{1 - e^{-\beta_2 t_0}} \right)^{\beta_3} + \varepsilon \quad (2)$$

其中, HD 指优势木平均高, β_1 为林分标准林龄时的优势高, β_2 为尺度参数, β_3 为形状参数, t 为林龄, t_0 为标准林龄(本文取 20 年), ε 为随机误差。

2.2 非线性混合效应模型

2.2.1 基本形式 一般来说, 单水平非线性混合效应模型的形式如(3)式:

$$\begin{cases} y_{ij} = f(\phi_{ij}, \nu_{ij}) + \varepsilon_{ij}, & i = 1, \dots, M, j = 1, \dots, n_i \\ \phi_{ij} = A_{ij}\beta + B_{ij}b_i \\ \varepsilon_{ij} \sim N(0, \sigma^2) & b_i \sim N(0, D) \\ b_i = \hat{DZ}_i^T (\hat{R}_i + \hat{Z}_i \hat{DZ}_i^T)^{-1} \hat{\varepsilon}_i \end{cases} \quad (3)$$

其中, y_{ij} 是第 i 个样地中第 j 次观测的因变量值, 本文指预测的样地林分优势木平均高。 M 是样地数, n_i 是第 i 个样地内观测的样地数。 f 是真实值, 是一个研究对象中具体参数向量 ϕ_{ij} 和变值向量 ν_{ij} 的可微函数, 本文指测量的样地林分优势木平均高。 ε_{ij} 是服从正态分布的误差项。 β 是 $(p \times 1)$ 维固定效应向量, b_i 是带有方差协方差矩阵 D 的 $(q \times 1)$ 维随机效应向量, A_{ij} 和 B_{ij} 是相应的设计矩阵。 \hat{D} 为样地间可变性的 $(q \times q)$ 维方差协方差矩阵 ($q =$ 随机效应参数个数), \hat{R}_i 是样地 i 的 $(k \times k)$ 维方差协方差矩阵, $\hat{\varepsilon}_i$ 是 $(k \times 1)$ 维残差向量, $\hat{\varepsilon}_{ij} = G_{ij} - f(\hat{\phi}_i, x_{ij})$, G_{ij} 为第 i 样地第 j 次测量的优势木平均高实测值。 $\hat{\phi}_i$ 仅仅包括固定参数向量, \hat{Z}_i 为参数 $\hat{\beta}$ 的 $(k \times q)$ 维矩阵。

2.2.2 确定模型结构 在构建一个混合模型之前, 需要确定以下 2 个结构^[9]:

(1) 误差效应方差协方差结构 (R_i)

对固定样地进行长期观测的优势木平均高数据通常都存在自相关和异方差问题。为了解决这两个问题, 林业上通常利用式(4)进行描述^[10-11]:

$$R_i = \sigma_i^2 \Psi_i^{0.5} \times \Gamma_i \times \Psi_i^{0.5} \quad (4)$$

其中, R_i 是样地内方差协方差矩阵, σ_i^2 为未知

的样地 i 的残差方差, Ψ_i 为描述样地内误差方差的异质性的对角矩阵, Γ_i 是描述误差效应自相关结构矩阵。

(2) 随机效应方差协方差结构 (D)

随机效应方差协方差结构反映了样地之间的变化性。本研究中所用的方差协方差结构有 3 种, 其中 m 为参数 β_1 的随机效应, n 为参数 β_2 的随机效应, p 为参数 β_3 的随机效应。选择 UN 方差协方差结构进行模拟。

2.3 模型精度比较

模型在模拟后要验证, 通常采用确定系数 (R^2)、均方误差 (MSE) 和平均绝对残差 ($|E|$) 3 个模型精度评价指标对模拟结果进行效果评价^[12]。具体公式见式(5~7)。

$$R^2 = r_{h_{ij}\hat{h}_{ij}}^2 \quad (5)$$

$$MSE = \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (h_{ij} - \hat{h}_{ij})^2}{m - 1} \quad (6)$$

$$|E| = \frac{\left| \sum_{i=1}^m \sum_{j=1}^{n_i} (h_{ij} - \hat{h}_{ij}) \right|}{m} \quad (7)$$

$r_{h_{ij}\hat{h}_{ij}}$ 为测量的优势木平均高 (h_{ij}) 和估计的优势木平均高 (\hat{h}_{ij}) 之间的相关系数, n_i 为样地内样地数量, m 为样地个数。

3 结果与分析

在模拟优势木平均高时, 首先确定模型的混合效应参数、误差效应方差协方差结构以及随机效应方差协方差结构, 然后利用 SAS 软件的 NLMIXED 模块对其中 10 块模拟样地数据进行参数的模拟, 把能够收敛的模型作为模拟结果。利用剩余的 6 块样地作为检验数据并计算 3 个精度评价指标。具体模拟及计算结果如下。

3.1 模型拟合结果

能够收敛的模型包括 4 种。其中, AIC (Akaike Information Criterion) 和 BIC (Bayesian Information Criterion) 值为评价模拟效果的指标。

从表 2 的结果来看, 4 个结果的 AIC 和 BIC 值都比传统的模型值小。这几个模型的模拟精度都随着混合效应参数个数的增加而提高。但是当考虑 β_1 为混合参数要比 β_2 作为混合参数的模拟效果要好的多, 这说明 β_1 极其不稳定, 对优势木平均高的模拟影响最大, 因此在模拟时首先考虑作为混合参数, 模拟结果 3 和结果 4 也说明了在包括 β_1 为混合参数时, 模型容易收敛, 并且效果也很好。为了进一步验

证模拟结果,对其余 6 块样地进行几种模拟结果的验证,见表 3。

表 2 混合效应模型模拟结果

模型形式	混合参数	AIC	BIC	参数估计值			误差效应方差协方差矩阵(R_i)	随机效应方差协方差矩阵(D)
				β_1	β_2	β_3		
模拟 1	β_1	151.3	155.1	14.096 5	0.175 1	1.501 7	$\sigma^2 = 0.090 55$	$\sigma_m^2 = 1.804 4$
模拟 2	β_2	339.7	343.6	15.130 0	0.120 3	1.197 5	$\sigma^2 = 0.346 6$	$\sigma_n^2 = 0.000 629$
模拟 3	β_1	85.6	91.0	14.054 9	0.179 9	1.513 2	$\sigma^2 = 0.049 19$	$\sigma_m^2 = 3.662 3$
	β_2							$\sigma_{mn} = -0.029 00$
模拟 4	β_1	77.6	83.0	14.087 9	0.174 0	1.476 6	$\sigma^2 = 0.046 83$	$\sigma_n^2 = 0.049 19$
	β_3							$\sigma_m^2 = 2.917 6$
最小二乘方法	无	440.1	452.4	14.093 4	0.173 3	1.480 3		$\sigma_{mp} = 0.188 7$ $\sigma_p^2 = 0.014 38$

表 3 模型验证结果

模型	均方误差 (MSE)	平均绝对残差 ($ E $)	确定系数 (R^2)
最小二乘	0.933 5	0.695 7	0.948 6
模拟 1	0.285 7	0.224 6	0.995 3
模拟 2	0.560 4	0.395 7	0.981 8
模拟 3	0.203 1	0.160 9	0.997 6
模拟 4	0.198 0	0.155 3	0.997 7

表 3 的验证结果说明,与传统最小二乘方法相比,混合模型方法提高了模型的精度,4 种混合模拟的结果也与表 2 的模拟结果一致。

3.2 考虑误差效应结构矩阵

3.2.1 时间序列相关性 本研究所用的断面积数据是连续 7 次观测,数据间存在着时间序列相关性。为了表达样地内的时间序列相关性,一阶自回归矩阵模型(AR(1))、一阶自回归与滑动平均模型相结合的矩阵模型(ARMA(1,1))及复合对称矩阵模型(CS)经常被用在森林生长收获模拟中^[11]。本文采用这 3 个自相关矩阵模型来表述林分断面积的自相

关结构。

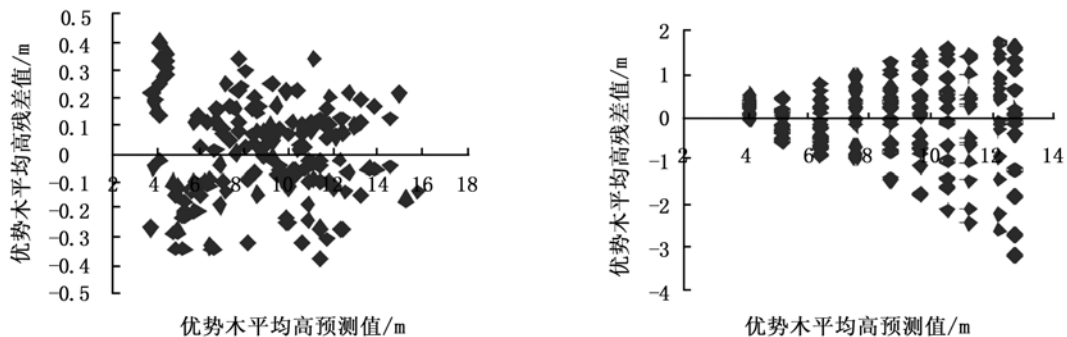
在 3.1 的 4 种混合效应模拟基础上,当考虑自相关性时,只有模拟结果 1 即 β_1 为混合参数时收敛,结果见表 4。

表 4 考虑时间序列相关性时混合效应模拟结果

模型形式	样本容量	时间序列相关结构	AIC	BIC
结构 1	105	不考虑	151.3	155.1
结构 2	105	AR(1)	67.2	71.8
结构 3	105	ARMA(1,1)	68.5	73.9
结构 4	105	CS		不收敛

通过表 4 的 AIC 和 BIC 两个指标,表明 Richards 方程适合选择 AR(1) 为误差方差协方差结构矩阵。

3.2.2 异方差 通常判断是否存在异方差最直观的方法就是利用断面积残差分布图,根据 3.1 的研究结果,选择模拟结果 1 与传统最小二乘方法进行异方差的比较。



a 混合效应模拟结果残差分布图 b 传统最小二乘方法残差分布图

图 1 模拟结果 1 混合效应模型结果与传统最小二乘模拟结果残差分布图

从图 1a 和图 1b 的对比图来看,采用最小二乘回归估计方法估计的优势木平均高与实际测量的优势木平均高的残差是随着预测值的增大而绝对值增大的,因此存在着明显的异方差问题。而采用混合效应模型方法则残差并不随着预测值的增大而增大,分布虽然并不完全均匀,但异方差问题明显减小,因此异方差的影响在本研究中并不考虑,即 $\Psi_i^{0.5} = I_i$ 。

3.3 初植密度对优势木平均高的影响

判断初植密度对优势木平均高生长是否有影响,可通过随机效应模型模拟的结果来决定,当 β_1 为随机参数时,结果为 I 区组 3 个样地的随机效应值都大于 0, II 区组除了 5 样地小于 0 外其余两个样地都大于 0,但其进行平均后值大于 0, III 区组除了样地 9 大于 0 外,其余两个样地都小于 0,对其进行平均后值接近于 0, IV 区和 V 区都小于 0。这说明在密度小的情况下,优势木平均高要大于平均值,在密度大的情况下优势木平均高小于平均值,而 III 区组代表了平均水平。这些说明初植密度对优势木平均高的生长有一定的影响,因此在模拟优势木平均高与林龄关系模型时可把初植密度作为哑变量考虑进去。最后确定的优势木平均高混合效应模型形式如下:

$$\left\{ \begin{array}{l} HD = \hat{H} + \varepsilon_i \\ \varepsilon_i \sim N(0, R_i) \\ \Gamma_i(\theta) = AR(1) \\ \Psi_i^{0.5} = I_i \\ R_i = \sigma^2 \Psi_i^{0.5} \times \Gamma_i(\theta) \times \Psi_i^{0.5} = \sigma^2 \Gamma_i(\theta) \\ \hat{H} = (\beta_1 + u_i + b_{11} * md_1 + b_{12} * md_2 + \\ \quad b_{13} * md_3 + b_{14} * md_4) * \left(\frac{1 - e^{-\beta_2 t}}{1 - e^{-\beta_2 t_0}} \right)^{\beta_3} \\ u_i \sim N(0, D) \\ i = 1, 2 \dots 16 \end{array} \right. \quad (8)$$

式中, md_1 表示初植密度为 I 区组时初植密度, md_2 表示初植密度为 II 区组初植密度, md_3 表示初植密度为 III 区组初植密度, md_4 表示初植密度为 IV 区组初植密度,当为 I 区组初植密度时, $md_1 = 1$, 而 md_2 、 md_3 和 md_4 为 0, 依次类推, 当初植密度为 V 区组时, md_1 、 md_2 、 md_3 和 md_4 都为 0。 b_{11} 、 b_{12} 、 b_{13} 和 b_{14} 为待估的固定效应参数, u_i 为待估的随机效应参数。

利用(8)式对优势木平均高数据进行拟合。结

果为 AIC 值等于 55.8、BIC 值等于 63.5, 比单纯考虑样地的随机效应、随机序列相关性拟合效果要好。最后对其余 6 块样地进行精度检验, 均方误差 MSE 为 0.2133、平均绝对残差 $|E|$ 为 0.1388、确定系数 R^2 为 0.9978。验证结果也说明当同时考虑多次观测的时间序列相关性、样地的随机效应及初植密度时, 拟合效果和模拟精度都有很大的提高。

4 结论与讨论

(1) 修改的 Richards 形式的优势木平均高与林龄关系的非线性混合效应模型的估计精度比传统的回归模型估计精度明显提高, 增加随机效应参数个数能够提高模型的估计精度。

(2) 通过对 Richards 方程的几种模拟精度比较, 可看出当考虑林分的标准年龄的优势高时避免了 β_1 的不稳定性, 因此提高了优势木平均高估计的准确性和稳定性。本研究所选择的模型参数中, β_1 是最灵活、变化最不稳定的参数, 在估计时最能够体现优势高的变化, 因此在拟合时应首先考虑其作为混合参数。

(3) 在模拟不同初植密度优势木平均高这样连续多次观测的纵向数据时, 与单独考虑样地的随机效应、时间序列相关性及初植密度相比, 同时考虑这 3 方面的因素不仅提高了模拟效果, 同时也能够反映不同密度对优势木平均高的影响。

(4) 由于非线性混合效应模型参数拟合方法的限制, 本次研究只考虑了样地间的随机效应而没有考虑区组间的随机效应, 因此, 还需进一步对样地效应进行研究。

参考文献:

- [1] 孟宪宇. 测树学[M]. 第2版, 北京: 中国林业出版社, 1994
- [2] Hall D B, Bailey R L. Modeling and prediction of forest growth variables based on multilevel nonlinear mixed models[J]. Forest Science, 2001, 47: 311 - 321
- [3] Fang Z, Bailey R L. Nonlinear mixed effects modeling for slash pine dominant height growth following intensive silvicultural treatments[J]. Forest Science, 2001, 47: 287 - 300
- [4] Calegario N, Daniels R F, Maestri R, et al. Modeling dominant height growth based on nonlinear mixed-effects model: a clonal *Eucalyptus* plantation case study[J]. Forest Ecology and Management, 2005, 204: 11 - 21
- [5] 李永慈, 唐守正. 用 Mixed 和 Nlmixed 过程建立混合生长模型[J]. 林业科学研究, 2004, 17(3): 279 - 283
- [6] Castedo F, Dieguez-Aranda U, Barrio M. A generalized height - diameter model including random components for radiata pine planta-

- tions in northwestern Spain [J]. *Forest Ecology and Management*, 2006, 229:202 – 213
- [7] 洪玲霞. 初植密度、间伐对杉木林分优势高生长过程的影响[J]. *林业科学研究*, 1997, 10(4):448 – 452
- [8] 章允清. 卫闽林场杉木人工林经验收获表的研制[J]. *福建林业科技*, 2006, 33(3):47 – 51
- [9] Dorado F C, Ulises D A, Marcos B A. *et al.* A generalized height-diameter model including random components for radiata pine plantations in northwestern Spain [J]. *Forest Ecology and Management*, 2006, 229:202 – 213
- [10] Fang Z, Bailey R L, Shiver B D. A multivariate simultaneous prediction system for stand growth and yield with fixed and random effects[J]. *Forest Science*, 2001, 47:550 – 562
- [11] Wang Y, Valerie M L, Thomas G B. Modelling and prediction of dominant height and site index of *Eucalyptus globules* plantations using a nonlinear mixed-effects model approach[J]. *Canadian Journal of Forest Research*, 2007, 37:1390 – 1403
- [12] Adame P, Rio M D, Canellas I. A mixed nonlinear height-diameter model for Pyrenean oak (*Quercus pyrenaica* Willd.) [J]. *Forest Ecology and Management*, 2008, 256:88 – 98