JOURNAL OF FOREST SCIENCE, *54*, 2008 (4): 176–182

Estimating the sample size for fitting taper equations

K. Kitikidou, G. Chatzilazarou

Department of Forestry and Management of the Environment and Natural Resources, Laboratory of Forest Biometry, Dimokritos University of Thrace, Orestiada, Greece

ABSTRACT: Much work has been done fitting taper equations to describe tree bole shapes, but few researchers have investigated how large the sample size should be. In this paper, a method that requires two variables that are linearly correlated was applied to determine the sample size for fitting taper equations. Two cases of sample size estimation were tested, based on the method mentioned above. In the first case, the sample size required is referred to the total number of diameters estimated in the sampled trees. In the second case, the sample size required is referred to the number of sampled trees. The analysis showed that both methods are efficient from a validity standpoint but the first method has the advantage of decreased cost, since it costs much more to incrementally sample another tree than it does to make another diameter measurement on an already sampled tree.

Keywords: sampling methods; tree shape; regression

Tree trunk diameter generally decreases from the base to the top. The way this reduction takes place determines the trunk form (Philip 1994). The comprehension of trunk form allows better estimates of trunk volume or biomass, better estimates for the kinds and quantities of various tree products, and better comprehension of competition and conditions of tree growth. One of the ways to describe tree bole shapes is by fitting taper equations. These are regression equations, linear or nonlinear, and they predict the diameter d_{h_i} at any tree height h_i.

The first step in fitting regression equations to data is the choice of a sufficiently large sample of representative observations. Almost every text or book on linear models addresses this question, but researchers who dealt with taper equation fitting usually determine the sample size arbitrarily, although a good sampling design for data collection is essential if we want to obtain an efficient, accurate and representative fit of the taper equation.

The aims of this study were to:

– present a method for calculating sample sizes of diameter and height measurements for taper curve fitting,
– examine if we can define the sample size as the number of diameter measurements (and not as the number of trees) and meet the error targets. In this way, we reduce the sampling cost, since a given number of observations could be measured on any number of individual trees.

REVIEW OF LITERATURE

There is a long tradition in the mathematical description of the diameter – height relationship. Beginning with the earliest taper equations by Höjer (1903) and Behre (1923, 1927), increasingly more complex functions have been introduced as methodologies and computational capabilities have developed. Nowadays, a wide variety of taper equations are described in forestry literature (Burkhart, Gregoire 1994).

However, many sampling aspects that should be accounted to guarantee a predefined error level at a minimal cost, which and how much data, are usually neglected in most papers on taper equations fitting. Indicatively are reported Goulding and Murray (1975), who used a sample of 1,267 trees, Max and

Burkhart (1976), who used a sample of 652 trees, Perez et al. (1990), who used a sample of 405 trees, and Kozak and Smith (1993), who used a sample of 603 trees. None of the researchers mentioned above apply any method of determination of the sample size.

However, the determination of the sample size in general has occupied the researchers. Concretely, Demaerschalk and Kozak (1974), based on Elfuing's (1952), Hoel's (1958), Laylock' s (1972) and Wynn's (1972) proposals, proposed a way of determining the sample size by using simple linear regression methods. If there is a clue that a linear relation between two variables is sufficiently strong, we can apply linear regression analysis in pre-sample (pilot-sample) data, estimate the arithmetic mean and variance of the independent variable, and finally estimate the sizes of the samples for each value of the independent variable, which depend on the acceptable error of the resulting dependent variable's estimate. The researcher predefines the acceptable error. From these sample sizes, the biggest is selected as the minimum required size of the final sample.

In case it is not possible to apply sampling methods for the independent variable, Demaerschalk and Kozak (1975) proposed an alternative solution, while Marshall and Demaerschalk (1986) extended their method, adding the possibility of analysis for an unequal cost of sampling per value of the independent variable. Elsiddig and Hetherington (1982) dealt with the determination of sample size by applying Demaerschalk's and Kozak's (1974) method in the construction of double entry volume tables. Elsiddig and Hetherington (1982) used the equation of simple linear regression, with the dependent variable being the total tree volume and the independent variable the square of breast height diameter multiplied by the total tree height.

Singh and Sedransk (1978) dealt with the determination of the required number of sampling points, in two-phase sampling, aiming at the application of multiple regression. Corona and Ferrara (1990, 1991) developed a method of sample size estimation, using the stand basal area increment as the dependent variable and the breast height diameter as the independent variable.

Cormier et al. (1992) examined how the sample size affected the standard error of the estimate in the least squares regression method in a taper model. Finally, Philip (1994) reported that in order to choose a minimum required sample size, we have to predefine a minimal acceptable error of the model that will describe a linear relation between two variables and estimate residual variance from the pre-sampling data.

MATERIALS AND METHODS

Material studied

The data used in this study come from measurements taken on 20 Hungarian oak (*Quercus conferta* Kit. or *Quercus frainetto* Ten.) trees. The trees were selected randomly from an area in Northern Greece (Cholomonda Chalkidiki), in order to cover the range of site qualities of the area (Kitikidou 2002). Each tree was measured for the diameter at stump height (30 cm above the ground), the diameter at 80-cm height above the ground, and the breast height diameter (1.3 m above the ground) with a measurement tape. After that, the diameters at two-metre intervals above the breast height diameter, i.e. at 3.3, 5.3, 7.3, … m above the ground were measured with Bitterlich's relascope. Thus, 136 diameters were measured in total. Finally, the total height of each tree was estimated with Bitterlich's relascope.

Sample size methods

Demaerschalk's and Kozak's method (1974), which was previously described, is widely used in sample size estimation. In this paper we attempt to test this method for general purposes of sample size estimation, specifically in taper equations. In order to apply Demaerschalk' s and Kozak's method, we need to find 2 variables that are linearly correlated. Simple linear regression has been used before in stem profile analysis (Gray 1956). In our data, we distinguished 2 cases: in the first case, the independent variable is the relative height and the dependent variable is the relative diameter of the pre-sample (pilot-sample) of 136 observations, which came from the measurements in 20 trees. In the second case, the independent variable is the breast height diameter of 20 trees and the dependent variable is the diameter at stump height, that means the pre-sample (pilot-sample) has 20 observations. The relative diameter is defined as d_{h_i}/D and the relative height is defined as h_i/H, where d_{h_i} is the diameter at height h_i, D is the breast height diameter and H is the total tree height.

At this point, we should clarify that linear regression was applied within a pooled data set across all the 20 trees of the pre-sample and not within each tree.

Then, simple linear regression analysis was applied between the independent variable Y and the

dependent variable X in each of the cases, using the statistical package SPSS (Norusis 2002; Kitikidou 2005). The criterion of reliability that indicates the acceptable error of the final estimate is the confidence interval of the means of predicted values for each value of the independent variable. The confidence interval width of mean Y_i, for specific X_i, is given by the equation:

$$w_i = 2t_{n-2,\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX}} \quad (1)$$

where:

$t_{n-2,\alpha/2}$ – value of t distribution for $(n-2)$ degrees of freedom and significance level α,

$\hat{\sigma}$ – standard error of the estimate,

n – final sample size,

$\bar{X}$ – mean of the independent variable distribution,

and

$$SSX = \sum_{i=1}^{n} (X_i - \bar{X})^2 = VarX\,(n-1) \quad (2)$$

where:

$VarX$ – variance of the independent variable distribution.

In order to find the t-value given on the right side of the equation (1), we must know the sample size n, which is the thing we are looking for, so approximations or iterative procedures are necessary (Freese 1956, 1962; Avery 1975). Demaershalk's and Kozak's method is based on confidence intervals for the mean in a simple linear regression from which the required sample size may be calculated to a predefined accuracy by simply resolving the confidence interval equation for N. Since the sample size for the i^{th} value of the independent variable N_i is implicitly present in the critical t-value, when the largest acceptable width of the confidence interval of estimated values W_i is predefined, the required sample size N_i for each X_i was calculated by the type:

$$W_i = 2t_{N_i-2,\alpha/2}\hat{\sigma}\sqrt{\frac{1}{N_i} + \frac{(X_i - \bar{X})^2}{SSX}} \quad (3)$$

where:

W_i – largest acceptable width of the confidence interval of estimated values, for each value of the independent variable, which is predefined and represents the acceptable error of estimate,

$t_{N_i-2,\alpha/2}$ – value of t distribution for (N_i-2) degrees of freedom and significance level α,

$\hat{\sigma}$ – standard error of the estimate, calculated from regression analysis in pre-sample data,

N_i – sample size for the i^{th} value of the independent variable,

and

$$SSX = VarX\,(N_i - 1) \quad (4)$$

By using types (3) and (4), the sizes of samples for each value of the independent variable were calculated and from them the largest size was selected as the final sample size (minimum required).

Basic regression hypothesis

The basic hypothesis that should be in effect in order to apply regression analysis, using the least squares method, is that the residuals should be normally distributed, with constant variance and zero mean. The violation of this hypothesis results in the confidence intervals and the tests of significance being invalid (Neter, Wasserman 1974; Neter et al. 1990). In order to test the regression residuals for their normality, variance and mean, the SPSS statistical package was used (Norusis 2002; Kitikidou 2005).

Kolmogorov-Smirnov test for one sample (Chakravarti et al. 1967) was used for normality testing. If (significance of Z) $\leq \alpha$, the distribution of a variable is far from normal, while if (significance of Z) $> \alpha$, the distribution of a variable is close to normal, for probability α.

For the homogeneity of variance test we applied Levene's test. The null hypothesis is:

H_0: The variables have homogeneous variance

and the alternative:

H_1: The variables do not have homogeneous variance.

We calculate the statistic

$$L = \frac{\left(\sum_{i=1}^{3} n_i - 3\right)\sum_{i=1}^{3}\left(n_i\left(\frac{\sum_{j=1}^{n_i}\left|Y_{ij} - T_i\right|}{n_i} - \frac{\sum_{i=1}^{3}\sum_{j=1}^{n_i}\left|Y_{ij} - T_i\right|}{\sum_{i=1}^{2} n_i}\right)^2\right)}{(k-1)\sum_{i=1}^{3}\sum_{j=1}^{n_i}\left(\left|Y_{ij} - T_i\right| - \frac{\sum_{j=1}^{n_i}\left|Y_{ij} - T_i\right|}{n_i}\right)^2}$$

where:

n_i – number of values of Y_i variable $(i = 1, 2, 3)$,

Y_{ij} – j^{th} value of the i^{th} variable $(j = 1, 2, \ldots, n_i)$,

T_i – trimmed mean of the i^{th} variable.

If (significance of L) $\leq \alpha$ we accept the H_1, while if (significance of L) $> \alpha$ we accept the H_0, for probability α. In order to test the homogeneity of variance of the regression residuals, we can check the scatter plot between the residuals and the values of the dependent variable, or better, we can check the scatter plot between the residuals and the predicted values (this is better because the residuals and the values

Table 1. Summary statistics for the two samples

Statistics	$\frac{d_{h_j}}{D}$	$\frac{h_i}{H}$	$d_{0.3}$ (m)	D (m)
Mean	0.762	0.382	0.158	0.131
Standard error	0.030	0.027	0.012	0.011
Median	0.833	0.294	0.150	0.125
Variance	0.124	0.097	0.003	0.002
Kurtosis	–0.937	–1.207	–0.594	–0.736
Skewness	–0.431	0.462	0.201	0.263
Range	1.347	0.965	0.200	0.165
Minimum	0.053	0.018	0.070	0.060
Maximum	1.400	0.983	0.270	0.225
Number of values	136	136	20	20

of the dependent variable are usually correlated, opposing to the residuals and the predicted values). When the points of the scatter plot give the impression that they are assembled in a thin horizontal strip around zero, without following any pattern, then the variance of the residuals is constant (Draper, Smith 1997).

RESULTS AND DISCUSSION

Sample statistics

Summary statistics for both samples (relative height – relative diameter, breast height diameter – stump height diameter) are given in Table 1. Each variable has a mean of 0.762, 0.382, 0.158 and 0.131 m, respectively. Their standard errors are 0.030, 0.027, 0.012 and 0.011, respectively.

Sample size methods

Figs. 1 and 2 show that linear regression is appropriate for both sample size methods. In the first case of sample size estimation, from the application of simple linear regression to the pre-sample data of 136 observations with the dependent variable being the relative diameter d_{h_i}/D where d_{h_i} is the diameter at height h_i and D the breast height diameter, and the independent variable being the relative height h_i/H where H is the total height, resulted the equation:

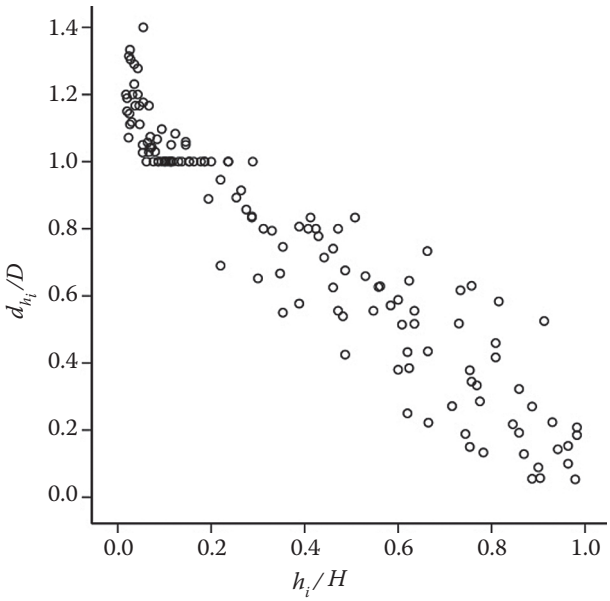

Fig. 1. Scatter plot of $d_{h_i}/D - h_i/H$

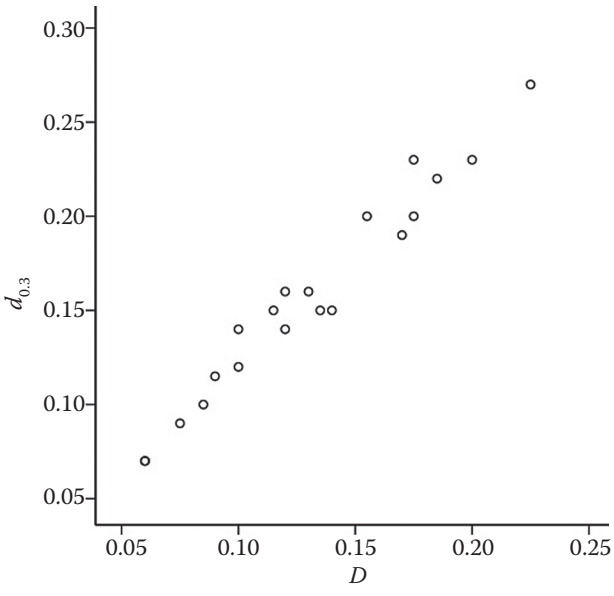

Fig. 2. Scatter plot of $d_{0.3} - D$

$$\frac{d_{h_i}}{D} = 1.171 - 1.073\frac{h_i}{H}$$

The equation was fitted to the data and resulted in a standard error of estimated values $\hat{\sigma} = 0.1123$ and an adjusted coefficient of determination $\overline{R^2} = 0.898$. Hypothesis tests for the regression coefficients resulted in large values of t (76.726 and –34.572 for the constant term and the coefficient of the independent variable, respectively), which results in significance of values less than 0.05 (0.0000 for both coefficients). Consequently, the two coefficients differ from zero ($P < 0.05$). The value of F from the analysis of variance was $F = 1{,}195.239$ ($P < 0.01$).

The acceptable error, which was predefined, was equal to 10% of the independent variable mean, that is:

$$W_i = 0.10\overline{X} = 0.10 \times 0.3817 = 0.03817$$

The most demanding value of the independent variable, as for the sample size, is $X_i = 0.9826$, which requires a sample size of 825 observations.

In the second case of sample size estimation, from the application of simple linear regression to the pre-sample data of 20 trees with the dependent variable being the diameter at stump height $d_{0.3}$ and the independent variable being the breast height diameter D, the following equation was obtained:

$$d_{0.3} = 1.202D$$

Regression analysis without a constant term was used because regression analysis with constant term resulted in a constant term value not different from zero ($P > 0.05$).

The equation was fitted to the data and resulted in a standard error of estimated values $\hat{\sigma} = 0.0111$ and an adjusted coefficient of determination $\overline{R^2} = 0.996$. Hypothesis tests for the regression coefficient resulted in a large value of t (66.856), which corresponds to $P < 0.05$. The value of F from the analysis of variance was $F = 4{,}469.734$, which corresponds to $P < 0.01$.

The acceptable error, which was predefined, was equal to 10% of the independent variable mean, that is:

$$W_i = 0.10\overline{X} = 0.10 \times 0.1308 = 0.01308$$

The most demanding value of the independent variable with respect to the sample size is $X_i = 0.2250$, which requires a sample size of 77 observations, that is 77 trees.

Basic regression hypothesis

Looking at Table 2, we see that for a 5% probability the residuals for both regressions approach the normal distribution (sig$Z$ = 0.348 > 0.05 and sig$Z$ = = 0.555 > 0.05). Also, the means of the residuals for both regressions are close to zero (Table 2). The significance of L is greater than the probability $\alpha = 0.05$ in the homogeneity of variance test; hence the residuals have homogeneous variance (Table 3). Figs. 3 and 4 show that the variances of residuals for both regressions are constant (the points of the graphics assemble in a thin horizontal strip around zero, with no obvious pattern).

CONCLUSIONS

For the estimation of a minimum required sample size for acceptable error of 10% of the independent variable mean, simple linear regression analysis was applied to data from a pre-sample of 20 trees. In the first case, the pre-sample had a total of 136 observations, which is the number of observed diameters in all 20 trees. The relative heights were used as the independent variable and the relative diameters as the dependent variable. The estimated minimum required size of the final sample calculated was 825 observations (825 diameters). In the second case, the

Table 2. Kolmogorov-Smirnov normality test

		Residuals (1st method)	Residuals (2nd method)
		$n = 136$	$n = 20$
Mean		0.0000	0.0006
Standard deviation		0.1119	0.0111
Most extreme differences	absolute	0.0800	0.1770
	positive	0.0800	0.1770
	negative	–0.0730	–0.0960
Kolmogorov-Smirnov Z		0.9330	0.7930
Asymptotic significance (2-tailed)		0.3480	0.5550

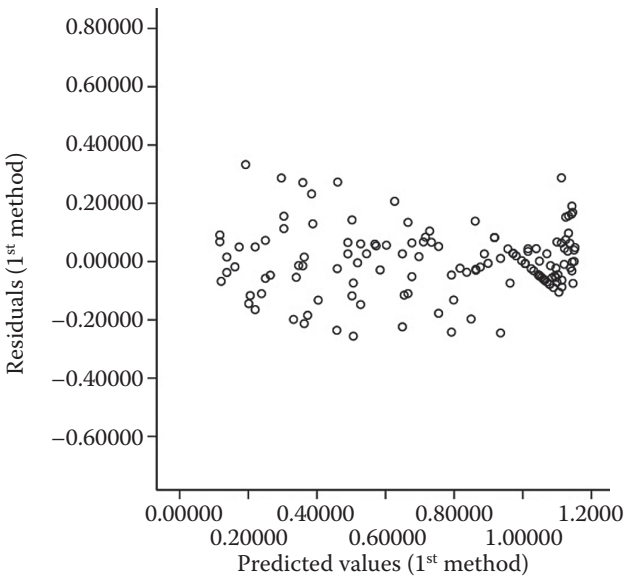

Fig. 3. Scatter plot for testing homogeneity of variance (1st regression line)

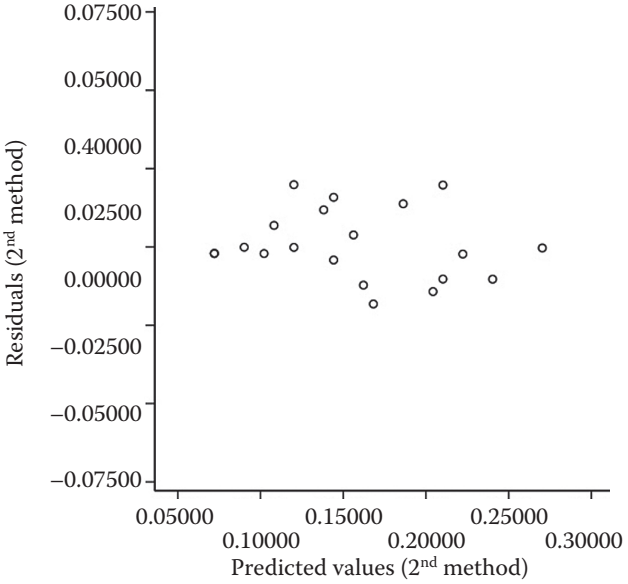

Fig. 4. Scatter plot for testing homogeneity of variance (2nd regression line)

Table 3. Levene's homogeneity of variance test

	Levene's statistic *L*	Significance
1st sample	0.205	0.652
2nd sample	1.509	0.235

pre-sample had a total of 20 observations. The breast height diameters were used as the independent variable and the diameters at stump height as the dependent variable. The minimum required size of the final sample calculated was 77 observations (77 trees). In both cases, the regression residuals were normally distributed, with constant variance and zero mean, so both methods are efficient from the validity standpoint. If we want to take into account both the cost of sampling and the precision, we must prefer the first case of a target number of individual diameters to observe, since we can measure a given number of diameters on any number of individual trees (less than the 77 trees that we found in the second case).

References

AVERY T., 1975. Natural Resources Measurements. 2nd Ed. New York, McGraw Hill Book Co.: 428.

BEHRE C., 1923. Preliminary notes on studies of tree form. Journal of Forestry, *21*: 507–511.

BEHRE C., 1927. Form-class Taper curves and volume tables and their application. Journal of Agricultural Research, *35*: 673–744.

BURKHART H., GREGOIRE T., 1994. Forest Biometrics. In: PATIL G., RAO C. (eds), Environmental Statistics (Handbook of Statistics). Amsterdam, Elsevier Science: 377–407.

CHAKRAVARTI I., LAHA R., ROY J., 1967. Handbook of Methods of Applied Statistics. Volume I. London, John Wiley and Sons, Inc.

CORMIER K., REICH R., CZAPLEWSKI R., BECHTOLD W., 1992. Evaluation of weighted regression and sample size developing a taper model for loblolly pine. Forest Ecology and Management, *53*: 65–76.

CORONA P., FERRARA A., 1990. Growth measuring techniques in forest assessment. Global Natural Resource Monitoring and Assessments: Preparing for the 21st century. In: Proceedings of the International Conference and Workshop, *3*: 1130–1132.

CORONA P., FERRARA A., 1991. Measuring techniques for assessing basal area increment of forest stands. Forest inventories in Europe with special reference to statistical methods. In: Proceedings of IUFRO Conference at Birmensdorf, Switzerland: 70–81.

DEMAERSCHALK J., KOZAK A., 1974. Suggestions and criteria for more effective regression sampling. Canadian Journal of Forest Research, *4*: 341–348.

DEMAERSCHALK J., KOZAK A., 1975. Suggestions and criteria for more effective regression sampling 2. Canadian Journal of Forest Research, *5*: 496–497.

DRAPER N., SMITH H., 1997. Applied Regression Analysis. New York, John Wiley and Sons, Inc.: 835.

ELFUING G., 1952. Optimum allocation in linear regression theory. Annals of Mathematical Statistics, *23*: 255–262.

ELSIDDIG E., HETHERINGTON J., 1982. The stem and branch volume of *Acacia nicotica* in the fung region in Su-

dan. Bangor N. Wales, University College of North Wales, Department of Forestry and Wood Science: 54.

FREESE F., 1956. Guidebook for Statistical Transients. USDA Forest Service, South Forest Experimental Station: 77.

FREESE F., 1962. Elementary Forest Sampling. USDA Forest Service, Agricultural Handbook No. 232: 91.

GOULDING C., MURRAY J., 1975. Polynomial taper equations that are compatible with tree volume equations. New Zealand Journal of Forest Science, *5*: 313–322.

GRAY H., 1956. The form and taper of forest tree stems. Oxford Imperial Forestry Institute, Paper No. 32: 79.

HOEL P., 1958. Efficiency problems in polynomial estimation. Annals of Mathematical Statistics, *29*: 1134–1150.

HOJER A., 1903. Growth of Scots pine and Norway spruce. Stockholm, Bilaga till. Loven, F.A. om vara barrskorar.

KITIKIDOU K., 2002. A study of the form of Hungarian oak trees (*Quercus conferta*) at Cholomonda Chalkidiki, Northern Greece. [Ph.D. Thesis.] Aristotelian University of Thessaloniki, Greece: 247. (In Greek)

KITIKIDOU K., 2005. Applied statistics with use of the SPSS statistical package. Thessaloniki, Tziola Publications: 288. (In Greek)

KOZAK A., SMITH J., 1993. Standards for evaluating taper estimating systems. The Forestry Chronicle, *69*: 438–444.

LAYLOCK P., 1972. Convex loss applied to design in regression problems. Journal of the Royal Statistical Association, *34*: 148–186.

MARSHALL P., DEMAERSCHALK J., 1986. A strategy for efficient sample selection in simple linear regression problems with unequal per unit sampling costs. The Forestry Chronicle, *62*: 16–19.

MAX T., BURKHART H., 1976. Segmented polynomial regression applied to taper equations. Forest Science, *22*: 283–289.

NETER J., WASSERMAN W., 1974. Applied Linear Statistical Models. Regression, Analysis of Variance and Experimental Designs. 3rd Ed. Boston, Richard D. Irwin, Inc.: 1184.

NETER J., WASSERMAN W., KUTNER M., 1990. Applied Linear Statistical Models. 3rd Ed. Homewood, Richard D. Irwin, Inc.: 1181.

NORUSIS M., 2002. SPSS 11.0 Guide to Data Analysis. Chicago, Prentice Hall: 637.

PEREZ D., BURKHART H., STIFF C., 1990. A variable-form taper function for *Pinus oocarpa* Schiede in central Honduras. Forest Science, *36*: 186–191.

PHILIP M., 1994. Measuring Trees and Forests. London, CAB International: 310.

SINGH B., SEDRANSK J., 1978. Sample size selection in regression analysis when there is nonresponse. Journal of the American Statistical Association, *73*: 362–365.

WYNN H., 1972. Results in the theory and construction of D-optimum experimental designs. Journal of the Royal Statistical Society, *34*: 133–147.

Received for publication October 12, 2007
Accepted after corrections February 12, 2008

Odhad vhodné velikosti vzorku ke stanovení tvarových křivek kmene

ABSTRAKT: Mnoho prací se zabývá hledáním vhodných tvarových křivek kmene, ale zatím bylo věnováno málo pozornosti otázce, jak velký vzorek měřených kmenů je k řešení této otázky dostatečný. V příspěvku je k nalezení vhodných rovnic použita metoda, vyžadující dvě proměnné, které jsou vzájemně lineárně korelovány. Testovány jsou potom dva soubory dat o různé velikosti. V prvním případě je velikost potřebného souboru dat vztažena k počtu průměrů měřených na jednotlivých kmenech. Ve druhém případě je velikost souboru vztažena k počtu kmenů. Analýza ukázala, že obě metody jsou vhodné z hlediska statistického hodnocení, ale první metoda je ekonomicky výhodnější. Je totiž snazší a levnější měřit další průměr na zvoleném kmeni, než měřit další kmen.

Klíčová slova: metody výběru; tvar kmene; regrese

Corresponding author:

Kitikidou Kyriaki, Ph.D., Dimokritos University of Thrace, Department of Forestry and Management of the Environment and Natural Resources, Laboratory of Forest Biometry, Anixis 14, Thessaloniki 54352, Greece
tel.: + 30 2310 914 743, e-mail: kkitikid@fmenr.duth.gr